IA

INGENIERÍA DE LA INTELIGENCIA ARTIFICIAL RESPONSABLE

AMPARO ALONSO BETANZOS, DANIEL PEÑA, PILAR PONCELA
Y CARLES SIERRA (EDITORES)

funcas

IA

INGENIERÍA DE LA INTELIGENCIA ARTIFICIAL RESPONSABLE

AMPARO ALONSO BETANZOS, DANIEL PEÑA, PILAR PONCELA Y CARLES SIERRA (EDITORES)

funcas

Impreso en España
Edita: Funcas
Caballero de Gracia, 28, 28013 - Madrid

ISBN impreso: 979-13-87770-16-7
ISBN digital: 979-13-87770-17-4
Depósito legal: M-5408-2026
Maquetación: Funcas

Contenido

Presentación

Este libro presenta de forma ampliada siete de las ponencias presentadas en la jornada que, con el título "El futuro de la inteligencia artificial", se organizó por los cuatro editores del libro en Funcas el 15 de octubre de 2025, y cuya grabación puede seguirse en la web de Funcas y en su canal de YouTube. Los trabajos incluidos en este libro abordan, desde distintos puntos de vista, el desarrollo futuro de la inteligencia artificial (IA), pero el título del libro se ha elegido para enfatizar un aspecto que consideramos fundamental en las contribuciones de los autores: la necesidad de un desarrollo responsable y socialmente beneficioso de la IA.

La IA se está convirtiendo cada vez más en una tecnología con profundas repercusiones socioeconómicas, que condiciona decisiones, comportamientos y oportunidades a escala social. Sin embargo, el debate público y técnico sigue aún centrado en el rendimiento (precisión, velocidad o eficiencia), y no tanto en las consecuencias, los valores y las formas de gobernanza que acompañan su despliegue. Este volumen parte de la idea de que el desarrollo futuro de la IA no depende únicamente de modelos más potentes, sino de nuestra capacidad colectiva para diseñar, evaluar y utilizar sistemas inteligentes de manera responsable, comprensible y socialmente legítima. Más allá de marcos éticos generales o normativos, quizás algo abstractos, los capítulos que componen este libro exploran cómo integrar la responsabilidad en el corazón mismo de la inteligencia artificial: en sus métricas, en sus arquitecturas de aprendizaje, en la formación de la ciudadanía, en la terminología que empleamos, en la comprensión de sus ciclos de evolución y, finalmente, en la gobernanza y la incorporación explícita de valores sociales en sistemas autónomos. Desde perspectivas complementarias, técnicas, epistemológicas, educativas y éticas, los autores ofrecen herramientas conceptuales y metodológicas para pensar una IA que no solo funcione bien, sino que también importe lo que hace, cómo lo hace y para quién lo hace.

En el capítulo I, "De la precisión a la responsabilidad: nuevas métricas para evaluar la IA", **Amparo Alonso Betanzos** ilustra con cuatro ejemplos la importancia de tener en cuenta las consecuencias sociales y económicas de la aplicación de la IA, ampliando el marco de evaluación para medir estas consecuencias e incluirlas junto a las medidas clásicas de eficacia de un sistema de decisión automática, como pre-

cisión, sensibilidad o eficiencia. Los ejemplos abordan cuatro temas que cubren un recorrido muy amplio y de gran actualidad. El primero es la identificación de usuarios tóxicos en redes sociales, un problema de importancia creciente que afecta desde el desarrollo de los adolescentes hasta la información difundida en redes sociales en las elecciones. El segundo, la mejora de los cada vez más usados sistemas de recomendación, buscando mayor explicabilidad y sostenibilidad. El tercero aborda las consecuencias de la priorización genética, que pretende identificar los genes con mayor probabilidad de asociación con procesos biológicos o patologías de interés. El cuarto ejemplo analiza el análisis de datos diádicos, que son aquellos que provienen de un par de entidades relacionadas, como dos usuarios de redes que interactúan entre sí. En estos datos, las predicciones sobre pares de entidades pueden tener consecuencias importantes en ámbitos sensibles que hay que tener en cuenta en los criterios de evaluación del sistema.

En el capítulo II, "Aprendizaje federado para una sociedad de máquinas", **Senén Barro**, **José Miguel Burés** y **Roberto Iglesias** proponen una inmersión profunda en lo que califican como un cambio de paradigma en el aprendizaje automático: la transición de modelos centralizados a estructuras de aprendizaje distribuidas y colaborativas. El capítulo describe cómo el Aprendizaje Federado (FL) permite resolver la tensión histórica entre la necesidad de grandes volúmenes de datos para el entrenamiento de modelos y el derecho irrenunciable a la privacidad de los usuarios. Los autores detallan la mecánica operativa del FL, donde el entrenamiento se desplaza hacia el "borde" (también llamado *edge computing*), permitiendo que dispositivos como teléfonos móviles o sensores médicos procesen información localmente. De esta forma, en lugar de transmitir datos sensibles a una nube central, solo se envían las actualizaciones de los parámetros del modelo local, que luego se agregan para mejorar el modelo global. El capítulo analiza con precisión los desafíos que aún limitan su aplicabilidad universal, como la heterogeneidad de los datos, la deriva de concepto (cómo gestionar modelos en entornos donde los datos cambian con el tiempo, también llamado *concept drift*), invalidando patrones previos, o la eficiencia en la comunicación. Finalmente, Barro y sus colaboradores subrayan que este enfoque no es solo una solución técnica que además aumenta la seguridad y la privacidad de los datos, sino la base de una "sociedad de máquinas". En este ecosistema, el aprendizaje deja de ser un proceso aislado para convertirse en un esfuerzo colectivo y democrático, donde la personalización del aprendizaje permite que la IA se adapte a las particularidades de cada individuo sin sacrificar el beneficio del conocimiento compartido.

En el capítulo III, "Alfabetización en IA: hacia un uso responsable y controlado de la tecnología", **Francisco Bellas** llama nuestra atención sobre la necesidad de alfabetización en IA para controlar y hacer un uso responsable de esta tecnología. No se trata de generar especialistas en IA, sino de educar a la población en general para que comprenda sus fundamentos, conozca sus derechos y deberes según el marco legal

europeo y desarrolle un pensamiento crítico sobre sus aplicaciones. Solamente así tendremos ciudadanos capacitados para usarla de manera responsable.

En el capítulo IV, titulado "Agentic AI y Multiagentic: ¿Estamos reinventando la rueda? Una reinterpretación epistemológica", su autor, **Vicent Botti,** aborda con rigor crítico la proliferación reciente de términos como IA agentiva o sistemas multiagentivos, argumentando que no constituyen una ruptura conceptual, sino una reelaboración tecnológica de paradigmas bien establecidos en el campo de los agentes autónomos y los sistemas multiagente. A través de un análisis detallado que recorre los fundamentos filosóficos, las definiciones clásicas y las arquitecturas consolidadas, el capítulo subraya la importancia de emplear una terminología precisa y de aprovechar el vasto acervo de investigación previa para evitar redundancias y fragmentación en el desarrollo de sistemas inteligentes. Esta postura se alinea de manera natural con los principios de la ingeniería de valores en inteligencia artificial, al promover un diseño que optimice recursos conceptuales y técnicos, garantice la interoperabilidad, y facilite la creación de soluciones robustas, éticas y sostenibles. Lejos de menospreciar los avances impulsados por los grandes modelos de lenguaje, el texto aboga por una integración consciente y fundamentada, donde la innovación se construya sobre cimientos sólidos y no sobre modas terminológicas. Así, el capítulo es una llamada al rigor epistemológico y a la continuidad disciplinar, invitando a la comunidad a enfocar sus esfuerzos en la verdadera evolución técnica y ética de la inteligencia artificial.

En el capítulo V, "Panarquía artificial: expansión y ciclos adaptativos de la IA", **David Martinez Rego** propone una forma diferente y especialmente sugerente de entender la evolución de la inteligencia artificial. Frente a la narrativa habitual que habla de avances lineales o de ciclos de euforia y decepción, el capítulo invita a mirar la IA como un sistema complejo, vivo y en constante adaptación, influido tanto por los avances tecnológicos como por factores sociales, económicos y culturales. Apoyándose en el concepto de panarquía, el autor describe cómo la IA evoluciona a través de ciclos interconectados de crecimiento, consolidación, crisis y reorganización. Esta perspectiva permite explicar por qué algunas tecnologías florecen de forma explosiva mientras otras quedan relegadas, por qué ciertas ideas resurgen décadas después con renovada fuerza, y por qué la innovación no ocurre al mismo ritmo en todos los niveles, desde la investigación básica hasta la adopción industrial y social. El capítulo recorre momentos clave de la historia reciente de la IA (desde los primeros enfoques estadísticos hasta el auge del aprendizaje profundo y los grandes modelos de lenguaje, LLMs– *Large Language Models*), mostrando que cada avance se apoya en equilibrios frágiles entre datos, computación, teoría y expectativas sociales. Lejos de presentar una visión triunfalista, el análisis subraya también las limitaciones, tensiones y riesgos que acompañan a cada ola tecnológica.

Con un enfoque divulgativo y reflexivo, Panarquía artificial ofrece al lector herramientas conceptuales para interpretar el presente y pensar el futuro de la IA con mayor profundidad. Más que responder a la pregunta de "¿qué puede hacer la IA?", el capítulo ayuda a comprender cómo y por qué evoluciona, y qué implica esa evolución para la toma de decisiones, la gobernanza tecnológica y el desarrollo responsable de sistemas inteligentes.

En el capítulo VI, titulado "Autonomía basada en valores: hacia una gobernanza responsable de los agentes IA", los autores **Holger Billhardt**, **Alberto Fernández**, **Andrés Holgado** y **Sascha Ossowski** abordan uno de los desafíos más urgentes y complejos de la inteligencia artificial contemporánea: cómo garantizar que los sistemas autónomos y multiagente actúen de forma alineada con los valores éticos de la sociedad que los integra. A través de un análisis riguroso que combina perspectivas técnicas, legales y éticas, el capítulo explora mecanismos de gobernanza para sistemas abiertos —como los basados en subastas para la gestión de intersecciones de tráfico— y presenta modelos formales para dotar a los agentes de una conciencia de valores que les permita razonar sobre dilemas morales y tomar decisiones responsables. Esta aproximación se enmarca de manera natural en los principios de la ingeniería de valores en inteligencia artificial, al ofrecer metodologías concretas para traducir valores abstractos —como la equidad, la eficiencia o la seguridad— en diseños computacionales verificables. Lejos de limitarse a la especulación teórica, el texto avanza hacia propuestas prácticas de aprendizaje de sistemas de valores a partir de preferencias observadas, integrando así la reflexión ética en el ciclo mismo de desarrollo de los agentes autónomos. De este modo, el capítulo no solo alerta sobre los riesgos de una autonomía desvinculada de los valores sociales, sino que proporciona herramientas fundamentales para construir sistemas de IA que sean técnicamente robustos, legalmente conformes y socialmente legítimos, estableciendo un puente indispensable entre la teoría ética y la práctica ingenieril en el panorama actual de la inteligencia artificial.

Finalmente, en el capítulo VII, "Máquinas que saben lo que importa: ingeniería de la conciencia de los valores sociales", **Carles Sierra** aborda el importante problema de desarrollar métodos de IA que incluyan representaciones computacionales de los valores, de forma que automáticamente sean tenidos en cuenta en la decisión o recomendación creada por el sistema de IA. En el capítulo se analizan métodos para incrustar valores en el sistema teniendo en cuenta los dilemas éticos y el contexto social. Se analizan estrategias de negociación basadas en valores y se describe un proceso de decisión donde un agente analiza el contexto y la conducta de la contraparte, evalúa los conflictos entre valores y aplica un criterio híbrido de decisión que satisface los valores del agente, pero dando cierto peso a los valores de la contraparte.

Los editores queremos agradecer a los autores de los capítulos su excelente disposición para cumplir los plazos previstos y realizar el esfuerzo de prepararlos pensando en un público no necesariamente especialista en el tema. Agradecemos también al director general de Funcas, Carlos Ocaña, el apoyo a esta iniciativa y a todas las actividades relacionadas con el *big data*, la IA y sus aplicaciones. Agradecemos a Cecilia y Esperanza su siempre eficaz labor de organización de las jornadas y a Myriam González y a su equipo la atenta y eficaz edición de este libro.

Amparo Alonso Betanzos, **Daniel Peña**, **Pilar Poncela** y **Carles Sierra**

Enero 2026

CAPÍTULO I

De la precisión a la responsabilidad: nuevas métricas para la inteligencia artificial

Amparo Alonso Betanzos*

La evaluación de los sistemas de inteligencia artificial ha estado tradicionalmente dominada por métricas de rendimiento técnico, como la precisión (proporción de predicciones positivas que son correctas), el *recall* o sensibilidad (proporción de casos positivos reales correctamente identificados) o el área bajo la curva (AUC, *Area Under Curve*, que mide la capacidad del modelo para discriminar entre clases a distintos umbrales de decisión). Sin embargo, este enfoque resulta hoy claramente insuficiente para capturar los efectos reales de la IA en contextos sociales complejos, como la discriminación de grupos, la falta de confianza de los usuarios o el elevado consumo energético de los algoritmos.

Este capítulo analiza cuatro casos de uso distintos con un objetivo común: mostrar por qué es imprescindible ampliar el marco de evaluación más allá del desempeño predictivo. Esta necesidad no es solo técnica —medir dimensiones como la sostenibilidad, la personalización o el impacto social—, sino también regulatoria. El AI Act europeo introduce la obligación de evaluar riesgos, efectos sociales y posibles daños que las métricas clásicas no reflejan. Sin métricas adecuadas, la propia aplicación de esta regulación se vuelve problemática: aquello que no puede medirse difícilmente puede gobernarse.

La ausencia de métricas integrales conlleva el riesgo de desplegar sistemas técnicamente excelentes pero éticamente deficientes, como modelos de reconocimiento facial con alta precisión y sesgos discriminatorios, o asistentes conversacionales fiables en apariencia pero opacos y propensos a alucinaciones. A través de los cuatro casos estudiados —la

* Este capítulo no hubiese sido posible sin las contribuciones de mis compañeros del grupo de investigación LIDIA (Laboratorio de I+D en IA) del CITIC de la Universidade da Coruña: Jorge Paz Ruza, Bertha Guijarro Berdiñas, Brais Cancela y Carlos Eiras Franco. Asimismo, agradezco la financiación de los proyectos nacionales PID2019-109238GB-C22 y PID2021-128045OA-I00 y de la Xunta de Galicia ED431C 2022/44, con fondos europeos ERDF.

predicción de toxicidad en plataformas digitales, las recomendaciones personalizadas con criterios de sostenibilidad, el uso ético del aprendizaje *Positive Unlabeled* para tratar datos sin etiquetar (cuando solo se conocen ejemplos positivos y el resto de los datos no están etiquetados) y la introducción de una métrica diádica (EAUC, *Excentricity Area Under the Curve*) para capturar sesgos como el de excentricidad—, el capítulo demuestra que lo que no se mide también importa. Justicia, transparencia y sostenibilidad ambiental deben incorporarse explícitamente a la evaluación, como condición necesaria para construir sistemas de IA que generen valor no solo técnico, sino también social.

Palabras clave: métricas en IA, sostenibilidad, sesgos, personalización.

1. INTRODUCCIÓN

La evaluación de modelos de inteligencia artificial (IA) ha estado tradicionalmente centrada en métricas de rendimiento, como la precisión, exactitud, error cuadrático medio, tasa de verdaderos positivos/negativos o área bajo la curva (AUC, que mide la capacidad del modelo para discriminar entre clases a distintos umbrales de decisión), entre otros. Estas métricas siguen siendo útiles, pero en la práctica actual resultan insuficientes. La rápida adopción de sistemas de IA en ámbitos críticos, como la educación, salud, finanzas, interacción social o gobernanza, ha puesto de manifiesto que el rendimiento técnico por sí solo no captura los efectos reales que los modelos producen en los usuarios, las instituciones y la sociedad (como podrían ser la discriminación, la falta de confianza o el consumo energético de los mismos, entre otros). De hecho, varios estudios han mostrado que modelos con altos niveles de precisión pueden simultáneamente perpetuar sesgos (Glickman and Sharot, 2025), producir decisiones opacas (Schilke and Reimann, 2025) o imponer costes ambientales y computacionales desproporcionados (Bolón-Canedo *et al.,* 2024), con importantes repercusiones sociales y económicas. Hoy en día, los modelos de IA no solo predicen, clasifican o recomiendan, sino que también interactúan con personas, influyen en comunidades, redistribuyen visibilidad, condicionan flujos de información y modifican comportamientos. Este cambio de escala y de función revela riesgos que no pueden evaluarse únicamente mediante las tradicionales métricas de rendimiento. Factores como la presencia de sesgos, la opacidad en la toma de decisiones, la posibilidad de generar o amplificar toxicidad, la falta de trazabilidad, la desigualdad en el acceso o las crecientes demandas computacionales y medioambientales son factores que deben integrarse en cualquier marco evaluativo contemporáneo. Paralelamente, el contexto regulatorio refuerza esa necesidad. La reciente aprobación del Reglamento de IA[1] en la Unión Europea exige que los sistemas sean verificables en dimensiones como equidad, transparencia, robustez, seguridad, sostenibilidad y responsabilidad social. Esto necesariamente obliga a ampliar la noción de evaluación para abarcar propiedades que no son estrictamente algorítmicas, sino también sociotécnicas: es decir, propiedades que emergen de la interacción entre modelos, datos, usuarios, instituciones y contextos de aplicación.

En este marco, la llamada IA Responsable (Goellner *et al.*, 2024) avanza hacia enfoques de evaluación multidimensionales que combinan rendimiento con criterios de impacto, sesgo, explicabilidad, responsabilidad y sostenibilidad. Estos nuevos enfoques buscan capturar aspectos tan diversos como la mitigación del daño, la explicabilidad operativa, el comportamiento emergente en sistemas interactivos, la adecuación contextual, la equidad en la distribución de resultados, la eficiencia energética, la reproducibilidad o la gobernanza del ciclo de vida del modelo. Este cambio de paradigma no implica descartar las métricas tradicionales, sino integrarlas en sistemas de evaluación más completos, capaces de reflejar dimensiones técnicas, sociales y éticas. Disponer de métricas

[1] Reglamento de IA: https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

robustas es, por lo tanto, esencial para garantizar que la IA se alinee con los valores y necesidades de la sociedad, minimizando riesgos y maximizando beneficios.

A pesar de estos avances, la brecha persiste. La mayoría de los estudios que proponen métricas para una IA ética se concentran en la equidad (Palumbo *et al.*, 2025) (paridad demográfica, probabilidades igualadas o impacto dispar) y tienden a limitarse en gran medida a entornos de datos tabulares o estructurados, con dificultades para generalizar a arquitecturas multimodales o complejas (por ejemplo, modelos que procesan texto, imágenes o grafos). Otras dimensiones críticas como la explicabilidad/transparencia (Deters *et al.*, 2025), la rendición de cuentas (*accountability*) (Raji *et al.*, 2020), la robustez, la gobernanza de datos, el bienestar social, la sostenibilidad[2] o la confianza del usuario carecen de métricas objetivas, ampliamente aceptadas y operacionalizables.

Esta carencia limita nuestra capacidad de construir sistemas de IA verdaderamente confiables. Sin métricas rigurosas y aplicables para la evaluación de la explicabilidad, el sesgo, la sostenibilidad, la robustez o el impacto social, los desarrolladores y reguladores carecen de herramientas para auditar y garantizar que los sistemas de IA cumplan con los estándares éticos y legales. Es urgente, por tanto, expandir la caja de herramientas de evaluación: definir, validar y adoptar un conjunto de métricas que capturen no solo qué hace un modelo, sino cómo y por qué lo hace, y qué consecuencias sociales puede implicar su comportamiento. Si aspiramos a que la IA no sea solo eficiente, sino también ética, humana y sostenible, debemos pasar de métricas puramente técnicas a sistemas de medición integrales que capturen esos pilares éticos. Este reto constituye tanto un desafío técnico como un imperativo social.

En este capítulo presentamos ejemplos concretos en los que hemos desarrollado o aplicado métricas que traducen principios éticos y legales en medidas cuantificables, útiles para el desarrollo y auditoría de sistemas de IA. En concreto, detallaremos cuatro casos: una medida de toxicidad predictiva para usuarios en redes sociales, un modelo de personalización explicable y sostenible para ser utilizado en un sistema de recomendación, un modelo sostenible que mejora la utilización de datos y la explicabilidad, así como el rendimiento de un sistema, y finalmente proponemos una métrica para sesgos en casos de regresión diádica. Si bien estos ejemplos no constituyen un marco completo multimodal de métricas, representan pasos significativos hacia la construcción de un sistema de evaluación integral para una IA responsable.

2. MODELO DE TOXICIDAD PREDICTIVA PARA USUARIOS DE REDES SOCIALES

Las plataformas digitales, como Reddit, Facebook o X, han intentado frenar la toxicidad en redes sociales utilizando un enfoque reactivo: detectar comentarios ofensivos y eliminarlos, con repercusiones como el bloqueo a los usuarios que los producen. Sin embargo,

[2] https://codecarbon.io/

la moderación tiene sus detractores y durante los últimos años ha habido cambios en el enfoque, en gran medida debido a que consume recursos que las compañías prefieren dedicar a otras cuestiones. Reddit, por ejemplo, en su último informe de transparencia de 2025[3] manifiesta que continúa usando una combinación de herramientas automatizadas y moderación humana para detectar y eliminar contenido que viola sus normas. Meta, por su parte, puso fin a su programa de verificación de hechos mediante terceros (*fact-checking* externo) en sus plataformas, como Facebook o Instagram, sustituyéndolo por un sistema tipo "Community Notes", similar al de X[4]. Esta última plataforma ha recortado sus recursos humanos en moderación, apostando por "moderar con la comunidad"[5]. Tras la eliminación de la moderación reactiva, varios estudios han detectado un aumento de denuncias en las plataformas por acoso, contenido violento y polarización[6].

En cualquier caso, este enfoque reactivo presenta limitaciones evidentes: llega tarde (al no evitar la propagación inicial de contenido nocivo), es costoso para las plataformas, genera desgaste psicológico en los moderadores y contribuye a que los usuarios más tóxicos migren a comunidades cada vez más extremas. En los últimos años, el problema de la toxicidad en redes sociales se ha agravado, en parte por el argumento, cada vez más utilizado por algunas plataformas, de que endurecer la moderación limitaría la "libertad de expresión" o frenaría la innovación. Sin embargo, medir la toxicidad de forma objetiva es muy complejo: depende del contexto cultural, del tono, de la intención y de matices lingüísticos que los sistemas actuales todavía no captan adecuadamente. De hecho, muchos modelos de moderación introducen sesgos sistemáticos que penalizan ciertas lenguas, estilos comunicativos o colectivos vulnerables, contribuyendo a desigualdades que, lejos de resolver el problema, lo amplifican. Además, los usuarios no son uniformes: su comportamiento varía notablemente según el espacio en el que participan, lo que demuestra que la toxicidad no es una propiedad fija del individuo, sino de la interacción entre usuario y comunidad.

Ante esta complejidad, necesitamos métricas transparentes, auditables y adaptables que permitan evaluar el fenómeno sin comprometer derechos ni reproducir discriminaciones. En este contexto, nuestra propuesta plantea un enfoque completamente distinto: predecir la toxicidad antes de que ocurra, bajo la asunción ya mencionada de que esta no depende únicamente ni del usuario ni de la comunidad, sino de la interacción entre ambas cuestiones. En lugar de identificar comentarios tóxicos ya publicados, el método anticipa si un determinado usuario va a comportarse de manera tóxica en una comunidad concreta (Paz-Ruza *et al*., 2025). Si esta predicción es correcta (y en nuestro caso de estudio se muestra que lo es en más del 80 % de las veces), las plataformas podrían evitar interacciones conflictivas antes de que tengan lugar, redu-

[3] https://redditinc.com/policies/transparency-report-january-to-june-2025-reddit

[4] https://techcrunch.com/2025/01/07/meta-drops-fact-checking-and-loosens-its-content-moderation-rules

[5] https://www.socialmediatoday.com/news/x-has-significantly-fewer-moderation-staff/714650/

[6] https://www.europapress.es/portaltic/socialmedia/noticia-meta-reduce-errores-moderacion-cambiar-sistema-aumenta-contenido-violento-facebook-20250530174039.html

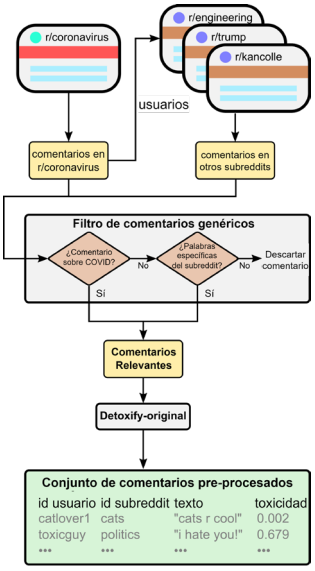
ciendo la exposición a contenido dañino en temas especialmente sensibles como la salud pública, y actuando preventivamente para reducir el riesgo sin recurrir a censuras indiscriminadas ni sobrecargar los sistemas de moderación.

Como caso de uso, hemos elaborado un modelo que hemos probado en la plataforma Reddit sobre comentarios del coronavirus. Para ello, recogemos los comentarios del subreddit coronavirus durante un período coincidente con los inicios de las campañas de vacunación; para cada usuario con comentarios en ese canal, añadimos hasta 1.000 comentarios de otros canales.

Posteriormente preprocesamos el conjunto extraído para, entre otras cuestiones, eliminar comentarios genéricos no útiles, y finalmente obtenemos un conjunto de 620.673 comentarios escritos por 25.893 usuarios en 331 subreddits diferentes. El proceso puede verse en la figura 1. Para cada comentario, nuestro objetivo es predecir su toxicidad, y para ello el método que usamos es el siguiente:

FIGURA 1

PROCESO DE ADQUISICIÓN, FILTRADO Y PREPROCESADO DE DATOS DE COMENTARIOS DE *REDDIT*



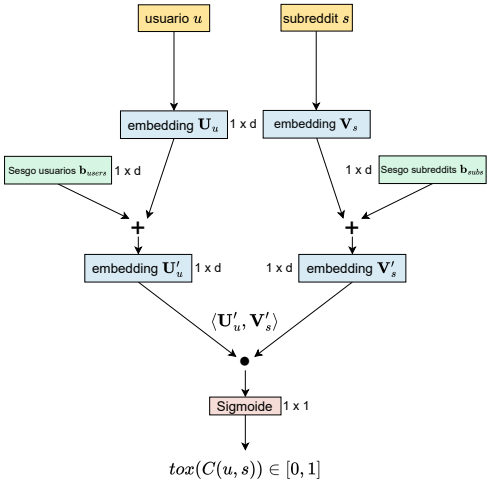
Fuente: Elaboración propia.

- Calculamos un valor [0,1] para la toxicidad utilizando un lenguaje preentrenado, Detoxify original (Hanu and [Unitory team], 2020), para medir la toxicidad de cada comentario.

- Utilizando un modelo de factorización matricial, cuyo esquema se muestra en la figura 2, aprendemos una representación latente de los usuarios y los subreddits, y su interacción predice la probabilidad de toxicidad futura.

FIGURA 2

TOPOLOGÍA DEL MODELO DE APRENDIZAJE AUTOMÁTICO PROPUESTO PARA PREDECIR LA TOXICIDAD DE CONVERSACIONES SOBRE COVID EN INTERACCIONES USUARIO-SUBREDDIT EN LA PLATAFORMA REDDIT. EL MODELO RECIBE LOS IDENTIFICADORES DEL USUARIO $\vec{u}$ Y SUBREDDIT $\vec{s}$, Y PREDICE LA TOXICIDAD ESPERADA $tox(C(\vec{u},\vec{s})) \in [0,1]$ DEL COMPORTAMIENTO DEL USUARIO EN EL SUBREDDIT. d ES EL NÚMERO DE CARACTERÍSTICAS LATENTES UTILIZADAS



Fuente: Elaboración propia.

Para poder predecir la toxicidad futura entre un usuario y un subreddit, es necesario que el modelo haya visto previamente ejemplos de ambos, lo que impide usar particiones tradicionales como la división fija de los datos de entrenamiento y prueba (*holdout*) o la validación cruzada, ya que podrían generar situaciones de ausencia de datos para entrenamiento (*cold start* o problema de inicio en frío). El método habitual en datos diádicos como los que nos ocupan, LOLI (*Leave One Last Item*) reserva para el conjunto de prueba la última interacción temporal (u, s) de cada usuario con dos o más interacciones. Sin embargo, este método es incompatible con nuestro problema, ya que nuestro conjunto de datos no dispone de una única etiqueta temporal por interacción y, además, LOLI está diseñado para tareas evaluadas mediante *rankings*. Por ello, hemos realizado una adaptación novedosa del método LOLI tradicional para tareas de clasificación binaria: seleccionar para cada usuario una de sus interacciones tóxicas y no tóxicas como test, garantizar que cada subreddit tenga ejemplos tanto en entrenamiento como en test y, si se desea, repetir el proceso para obtener un conjunto de validación. De este modo, se construyen particiones equilibradas y realistas para entrenar y evaluar el modelo.

Para evaluar la capacidad predictiva del modelo, utilizamos métricas clásicas de clasificación binaria: sensibilidad, especificidad y la media geométrica (G.Mean) de ambas clases. No se empleó la exactitud, por ser inapropiada en conjuntos desbalanceados como el caso concreto que comentamos, ni el AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*), ya que tiende a ser excesivamente optimista en escenarios muy desequilibrados (Imani *et al.*, 2026) al no reflejar directamente el rendimiento sobre la clase minoritaria. Tampoco se utilizó el F1-Score, definido como la media armónica entre precisión y recall, dado que en escenarios muy desbalanceados puede ser inadecuado al no considerar los verdaderos negativos ni reflejar el impacto global de los errores. La media geométrica resulta más apropiada cuando ambas clases son igualmente relevantes, como ocurre en este caso: tanto bloquear usuarios no tóxicos como permitir la interacción de usuarios tóxicos empeoran la idoneidad de actuación en la plataforma.

Por otra parte, la propuesta introduce un enfoque predictivo novedoso, distinto de los métodos tradicionales basados en detectar y reaccionar, por lo que no existen técnicas directamente comparables en el estado del arte. Por ello, se definieron tres líneas de base de complejidad creciente:

- NON, que predice todas las interacciones como no tóxicas (imitando el comportamiento de no anticipar nada);
- RND, que asigna toxicidad de forma aleatoria según la proporción observada en el conjunto de entrenamiento ($p \approx 0{,}1$);
- USR, que predice la toxicidad de un usuario según la frecuencia de interacciones tóxicas en su historial.

Como podemos ver en la tabla 1, el modelo, que propone un nuevo enfoque predictivo para combatir la toxicidad más eficiente y menos dañino que la moderación reactiva,

TABLA 1

EVALUACIÓN EN EL CONJUNTO DE PRUEBA DE NUESTRO MODELO (MDL), JUNTO CON LOS MODELOS BASE. RESULTADOS DE SENSIBILIDAD (QUÉ PROPORCIÓN DE CASOS TÓXICOS DETECTA EL MODELO), ESPECIFICIDAD (QUÉ PROPORCIÓN DE CASOS NO TÓXICOS CLASIFICA CORRECTAMENTE) Y MEDIA GEOMÉTRICA. LOS MEJORES RESULTADOS SE MARCAN EN AZUL, Y EL SEGUNDO MEJOR RESULTADO ESTÁ EN NEGRITA EN LA TABLA

	Sensibilidad	*Especificidad*	*Media geométrica*
NON	0.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
RND	0.100 ± 0.011	0.907 ± 0.014	0.301 ± 0.010
USR	**0.241 ± 0.012**	**0.910 ± 0.004**	**0.468 ± 0.010**
MDL	0.849 ± 0.023	0.810 ± 0.017	0.829 ± 0.018

ofrece un rendimiento elevado en un escenario difícil y muy desbalanceado, detectando cuatro veces mejor la toxicidad futura que las estrategias basadas en comportamiento pasado del usuario, evitando el sesgo de predecir todo como el caso general "no tóxico".

Este caso de uso pone de manifiesto que anticipar la toxicidad no es únicamente un problema de modelado, sino también un problema de evaluación adecuada. La capacidad de predecir interacciones tóxicas futuras depende críticamente de cómo construimos los conjuntos de datos, de qué métricas utilizamos y de qué tipo de errores consideramos aceptables. Medir correctamente la toxicidad —y no solo detectarla *a posteriori*— permite diseñar intervenciones más proporcionales, menos punitivas y potencialmente más justas para los usuarios, reduciendo tanto el daño causado por contenidos tóxicos como los efectos colaterales de una moderación excesivamente reactiva. Este ejemplo ilustra, por tanto, que ampliar el marco de evaluación es una condición necesaria para anticipar riesgos sociales en plataformas digitales.

3. MODELO PARA LA EXPLICABILIDAD SOSTENIBLE Y PERSONALIZADA EN SISTEMAS DE RECOMENDACIÓN

La explicabilidad y la sostenibilidad son otros dos de los pilares de la IA responsable. Los sistemas inteligentes deben poder explicar sus razonamientos y recomendaciones para poder crear confianza en los usuarios. Y deben hacerlo en lo posible sin añadir costes computacionales, cada vez más altos dada la creciente complejidad de estos sistemas. En este caso de uso veremos esta problemática dentro de un tipo concreto de sistemas, que son los sistemas recomendadores.

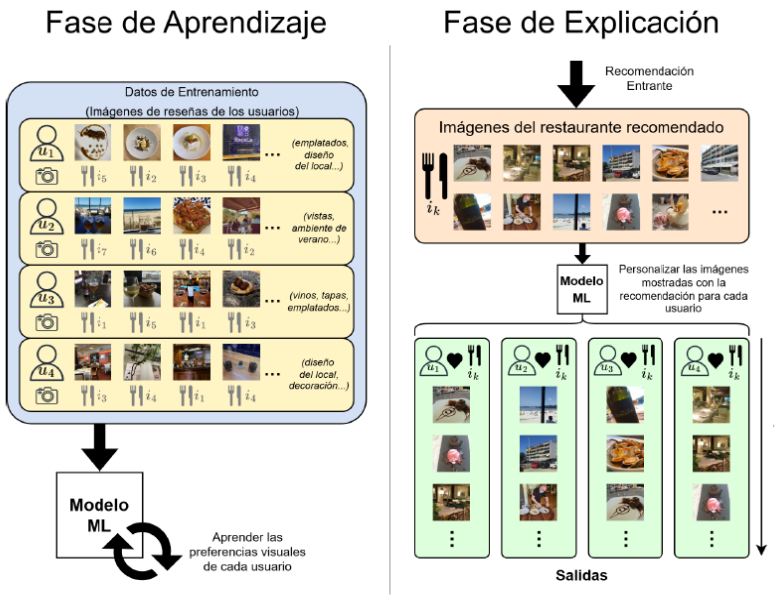
La creciente adopción de sistemas de recomendación en los que se usa IA ha traído enormes beneficios en cuanto a personalización, escalabilidad y eficiencia. Sin embargo, la opacidad de muchos de estos sistemas sigue siendo una preocupación estructural. Por ello, a medida que los recomendadores automatizados influyen en las decisiones cotidianas de personas (qué productos comprar, qué contenidos consumir, qué decisiones culturales o sociales tomar), la necesidad de explicabilidad, trazabilidad y confianza del usuario se vuelve crítica. Esta situación evidencia también la carencia de métricas de transparencia y explicabilidad adaptadas a escenarios reales. Existe, por tanto, una brecha entre lo que optimizan los modelos y aquello que realmente necesitan los usuarios para confiar en un sistema de IA.

En este contexto, nuestro trabajo no introduce una nueva métrica, sino un método que evidencia las limitaciones de las métricas existentes y que, además, muestra que es posible construir sistemas explicables más alineados con el usuario sin aumentar la complejidad del modelo ni el coste computacional y energético. El método propuesto permite generar explicaciones más personalizadas —adaptadas al usuario y al ítem—, ya que los usuarios ya no solo buscan que las recomendaciones que reciben sean personalizadas, sino que también exista cierta explicación que justifique dicha recomendación. Una vía

para lograr esta explicabilidad es el uso de contenido visual, por ejemplo, fotografías de productos, imágenes de lugares, fotos subidas por usuarios, etc., que es bastante común y natural en la percepción humana. Este enfoque natural promueve explicaciones que no son generadas artificialmente ni dependen de metadatos adicionales, lo que puede aumentar la confianza del usuario. La aproximación que aquí interesa es utilizar imágenes subidas a la plataforma en concreto por los propios usuarios como medio de explicación. Este método es intrínsecamente más orgánico y confiable que generar imágenes sintéticas o depender de textos auxiliares, ya que las explicaciones visuales son más representativas de las características del ítem (ver figura 3).

FIGURA 3

ESQUEMA GENERAL DE LA UTILIZACIÓN DE IMÁGENES SUBIDAS POR LOS USUARIOS PARA UNA EXPLICABILIDAD BASADA EN LO VISUAL Y RECOMENDACIONES PERSONALIZADAS. LOS USUARIOS SUBEN IMÁGENES A UNA PLATAFORMA PARA JUSTIFICAR SUS EXPERIENCIAS CON UN ÍTEM. SE CONSIDERA QUE LAS CARACTERÍSTICAS DE ESTAS IMÁGENES CONSTITUYEN BUENAS EXPLICACIONES PARA ESE USUARIO, Y LA IDEA ES QUE UN MODELO DE APRENDIZAJE PUEDE APRENDER ESTAS CARACTERÍSTICAS PARA CADA USUARIO (PARTE IZQUIERDA DE LA FIGURA). LUEGO, LA APARIENCIA DE CUALQUIER RECOMENDACIÓN ENTRANTE PUEDE PERSONALIZARSE (EN ESTE CASO, CON UNA IMAGEN DEL RESTAURANTE RECOMENDADO) UTILIZANDO LA IMAGEN YA EXISTENTE EN LA PLATAFORMA QUE MEJOR REFLEJE LAS PREFERENCIAS EXPLICATIVAS QUE SE HAN APRENDIDO DEL USUARIO; LA IMAGEN ELEGIDA COMO EXPLICACIÓN VARIARÁ ENTRE USUARIOS SEGÚN SUS PREFERENCIAS (LADO DERECHO)



Fuente: Elaboración propia.

Sin embargo, emplear imágenes subidas por el usuario a una plataforma plantea ciertos desafíos importantes, como es el caso de la dispersión de los datos —mientras que las explicaciones textuales pueden ser compartidas por varios usuarios o ítems, las imágenes son únicas para ese par (usuario, ítem)—, lo que hace que el modelo tenga más dificultad para aprender, y además tengamos el problema de que los datos no específicamente subidos por el usuario son normalmente tratados como datos negativos (algo que claramente no es cierto necesariamente, ya que a un usuario determinado puede gustarle una imagen que haya subido otro usuario diferente sobre el mismo ítem o incluso sobre otro diferente). Estas cuestiones tienen además mayor influencia en los modelos que hasta ahora han sido el estado del arte (Paz-Ruza *et al.*, 2022; Diez *et al.*, 2020), en los que el problema se aborda como una tarea de clasificación de autoría, es decir, prediciendo qué imagen de un ítem tiene más probabilidades de haber sido subida por el usuario específico. No obstante, este enfoque de modelar una tarea de *ranking* como una clasificación binaria es subóptimo y conlleva altos costes computacionales y modelos voluminosos. Al centrarse en un *ranking* eficiente y no en modelos complejos de clasificación, nuestro modelo BRIE (Paz-Ruza *et al.*, 2024) consigue reducciones sustanciales de coste computacional y energético, una dimensión que rara vez está integrada en los marcos métricos tradicionales, pese a ser esencial para una IA sostenible (Schwartz *et al.*, 2020; Bolón-Canedo *et al.*, 2024). Este cambio de perspectiva es importante porque revela un desajuste profundo: muchas métricas hoy utilizadas para evaluar explicabilidad no capturan ni la relevancia subjetiva de las explicaciones, ni su adecuación al usuario, ni su impacto en la confianza en el sistema.

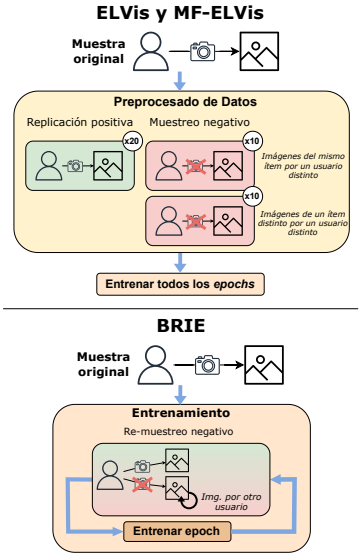
Como mencionamos anteriormente, los enfoques del estado del arte, llamados ELVis (Diez *et al.,* 2020) y MF-ELVis (Paz-Ruza *et al.,* 2022), abordan la selección de explicaciones visuales formulando el problema como una tarea de clasificación binaria. En estos métodos, lo que ocurre en términos generales es que el sistema aprende a distinguir entre explicaciones "buenas" y"malas" para, de forma indirecta, aproximar el problema real de ordenar explicaciones según su adecuación para un usuario concreto. Nuestra propuesta adopta una estrategia más directa basada en aprendizaje por *ranking*, utilizando el algoritmo Bayesian Pairwise Ranking (BPR). Este método está diseñado específicamente para aprender funciones de ordenación, ajustando el modelo para que las explicaciones consideradas como adecuadas aparezcan sistemáticamente por encima de las menos relevantes.

Para ello, BRIE asume que las imágenes subidas por un usuario reflejan de forma consistente los aspectos que este considera relevantes al justificar sus opiniones. Por tanto, todas las imágenes aportadas por un usuario se tratan como ejemplos positivos, bajo la hipótesis de que existe coherencia en los criterios visuales que cada usuario emplea. Como ejemplos negativos, se consideran imágenes no subidas por dicho usuario, una simplificación habitual en escenarios donde no se dispone de ejemplos negativos claramente definidos, un aspecto que en la actualidad estamos mejorando utilizando aproximaciones que nos ayuden a una identificación más precisa de los ejemplos negativos

para cada usuario concreto. BRIE introduce un mecanismo de muestreo más sencillo y flexible: en cada iteración de entrenamiento se generan nuevos ejemplos negativos de forma aleatoria, sin necesidad de preseleccionarlos ni almacenarlos, como se muestra en la figura 4. Esto reduce la complejidad del proceso de aprendizaje y permite un entrenamiento más eficiente, manteniendo al mismo tiempo la capacidad del modelo para aprender preferencias visuales personalizadas.

FIGURA 4

SELECCIÓN DE MUESTRAS NEGATIVAS DE LOS MÉTODOS DEL ESTADO DEL ARTE, LLAMADOS ELVIS Y MF-ELVIS (ARRIBA) Y DEL MÉTODO PROPUESTO, BRIE (ABAJO), USADAS PARA CADA MUESTRA (u, i, p) DEL CONJUNTO DE ENTRENAMIENTO $\mathcal{D}$

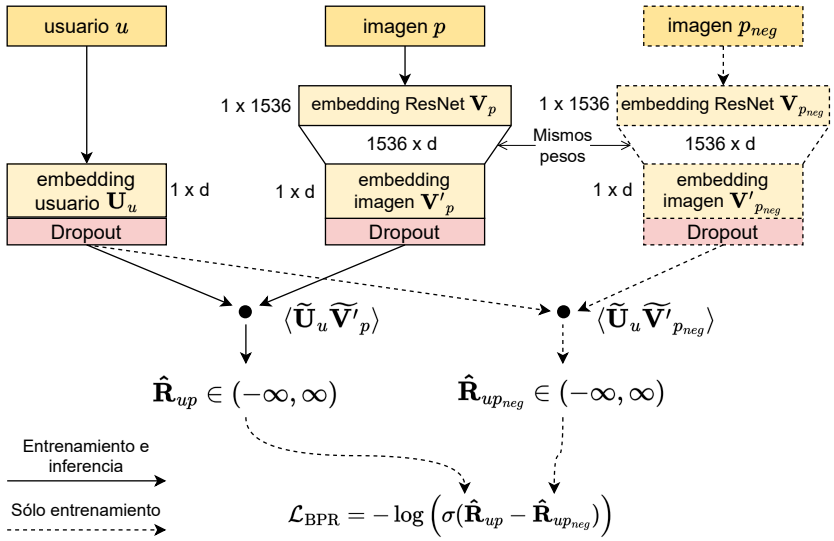


Fuente: Elaboración propia.

Como se muestra en la figura 5, BRIE mantiene la estructura general de modelos anteriores como los ya mencionados ELVis y MF-ELVis, pero introduce mejoras en la forma en que se usan los datos de entrenamiento. El modelo aprende representaciones (*embeddings*) tanto de los usuarios como de las imágenes. Durante el entrenamiento, BRIE recibe un usuario, una imagen subida por ese usuario (considerada adecuada) y otra imagen negativa (considerada poco adecuada). El objetivo es que el sistema aprenda a dar una puntuación más alta a la imagen correcta que a la negativa para ese usuario. En la fase de uso real (inferencia), el modelo solo necesita un usuario y una imagen y devuelve una puntuación que indica lo adecuada que es esa imagen como explicación para dicho usuario. Este enfoque basado en *ranking* simplifica el modelo y evita tratar el problema como una clasificación binaria.

FIGURA 5

TOPOLOGÍA DE BRIE. EN EL PROCESO DE ENTRENAMIENTO, RECIBE EL ID DE USUARIO u, UNA IMAGEN p TOMADA POR u, Y OTRA p_{neg} CONSIDERADA UNA EXPLICACIÓN INADECUADA PARA u, Y DEBE MAXIMIZAR LA DIFERENCIA ENTRE SUS AUTORÍAS PREDICHAS ($\hat{\mathrm{R}}_{up} - \hat{\mathrm{R}}_{up_{neg}}$) PARA MINIMIZAR LA PÉRDIDA $\mathcal{L}_{BPR}$. EN INFERENCIA, RECIBE EL ID DE u Y UNA IMAGEN p, Y CALCULA LA AUTORÍA PREDICHA $\hat{\mathrm{R}}_{up}$, I.E., LA IDONEIDAD DE p PARA EXPLICAR A u UNA RECOMENDACIÓN DEL ÍTEM



Fuente: Elaboración propia.

Para comprobar el rendimiento computacional del modelo, usamos cuatro conjuntos de datos con reseñas de diferentes restaurantes con imágenes tomadas por usuarios de la plataforma TripAdvisor. Los conjuntos de datos pertenecen a distintas ciudades de diferente tamaño u contexto cultural y geográfico (Gijón, Barcelona, Nueva York y Londres) (ver tabla 2), comparando nuestra aproximación con los dos modelos estado del arte, ELVis y MF-ELVis, y con dos modelos de referencia sencillos, el modelo aleatorio RND (Random) —que asigna una preferencia aleatoria a cada par (usuario,imagen)— y un modelo basado en distancias, CNT, que estima la preferencia

TABLA 2

CARACTERÍSTICAS BÁSICAS DE CADA CONJUNTO DE DATOS

Ciudad	*Usuarios*	*Restaurantes*	*Imágenes*	*Imágenes/usuario*
Gijón	5.139	598	18.679	3,64
Nueva York	61.019	7.588	231.141	3,79
Londres	134.816	13.888	479.798	3,56

del par (usuario,imagen) como la distancia negativa entre el embedding ResNetv2 Vp de la imagen p y el centroide de los embeddings ResNetv2 de las imágenes tomadas por el usuario u. Como métricas, usamos Recall@k, NDCG@k y AUC en el *ranking* f(u, i) creado por el modelo para cada caso de prueba, e informamos de las métricas medias (MRecall@k, MNDCG@k, MAUC) de todos los casos de prueba para un conjunto de datos y un modelo (tabla 3).

TABLA 3

RENDIMIENTO DE LOS MODELOS EN CADA CONJUNTO. EL MEJOR Y SEGUNDO MEJOR RESULTADO SE RESALTAN EN NEGRITA Y SUBRAYADO, RESPECTIVAMENTE

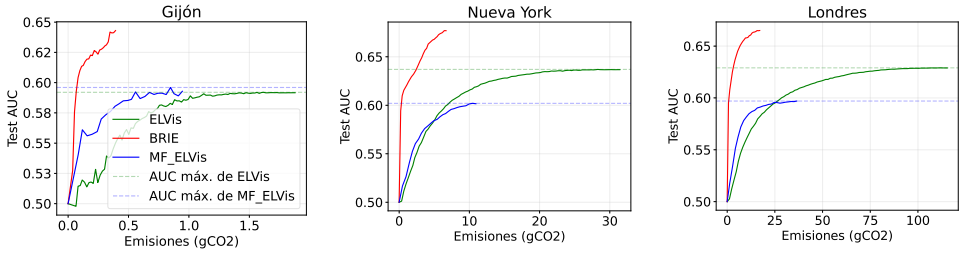
		Métricas		
Ciudad	*Modelo*	*MRecall@10*	*MNDCG@10*	*MAUC*
*Gijón	RND	0.373	0.185	0.487
	CNT	0.464	0.218	0.546
	ELVis	0.521	0.262	**0.596**
	MF-ELVis	**0.538**	**0.285**	0.592
	BRIE	0.607	0.333	0.643
*Nueva York	RND	0.374	0.168	0.502
	CNT	0.431	0.217	0.563
	ELVis	**0.553**	**0.304**	**0.637**
	MF-ELVis	0.516	0.276	0.602
	BRIE	0.598	0.341	0.677
*Londres	RND	0.342	0.155	0.500
	CNT	0.400	0.200	0.562
	ELVis	0.530	**0.293**	**0.629**
	MF-ELVis	**0.531**	0.267	0.597
	BRIE	0.563	0.318	0.665

El modelo BRIE no solo mejora el rendimiento respecto a los métodos anteriores del estado del arte, sino que también reduce de forma significativa el impacto medioambiental. Durante el entrenamiento, BRIE emite aproximadamente un 75 % menos de CO_2 que ELVis y un 50 % menos que MF-ELVis, manteniendo además mejores resultados en todas las métricas evaluadas (Recall, NDCG y AUC) en los tres conjuntos de datos analizados (Gijón, Nueva York y Londres), como vemos en la figura 6.

Los experimentos muestran que BRIE logra un mejor compromiso entre rendimiento y sostenibilidad, de modo que, independientemente del objetivo de reducción de emi-

FIGURA 6

COMPARACIÓN DEL AUC EN TEST VS. EMISIONES DE CO_2 DURANTE EL ENTRENAMIENTO PARA CADA MODELO Y CIUDAD

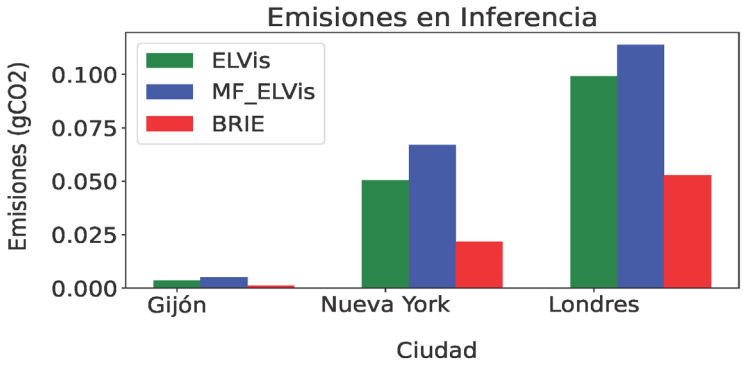


Fuente: Elaboración propia.

siones que se persiga, es posible alcanzarlo sin sacrificar calidad predictiva. En fase de inferencia, BRIE también resulta más eficiente, con una reducción cercana al 50 % en las emisiones de CO_2 frente a ELVis y MF-ELVis en escenarios de uso real, como podemos ver en la figura 7.

FIGURA 7

COMPARACIÓN DE EMISIONES DE CO_2 EN INFERENCIA DE LA PARTICIÓN DE TEST DE CADA CONJUNTO; PROMEDIO SOBRE 50 EJECUCIONES Y CON ACELERACIÓN GPU



Fuente: Elaboración propia.

Esta mejora se debe principalmente a una política de entrenamiento más adecuada y a un uso más eficiente de los datos, evitando arquitecturas innecesariamente complejas. Aunque los experimentos se han realizado con aceleración GPU, lo que mitiga parcialmente el coste de los modelos más grandes, en escenarios basados únicamente en CPU la ventaja energética de BRIE sería aún mayor.

Como vemos, es posible diseñar modelos explicativos eficaces para sistemas de recomendación basados en imágenes de usuarios, que optimicen rendimiento y eficiencia de datos sin sacrificar sostenibilidad. Resaltamos la necesidad de nuevas métricas que evalúen no solo precisión, sino también eficiencia energética y sostenibilidad ambiental de los modelos. Nuestro enfoque demuestra que es posible avanzar en esas direcciones incluso sin métricas nuevas, pero también evidencia la urgencia de desarrollarlas. Sin ellas, la evaluación de sistemas explicables seguirá siendo parcial y no reflejará los valores —personalización, sostenibilidad, confianza— que queremos promover en la IA contemporánea.

4. UN MODELO PARA MEJORAR EL APRENDIZAJE AUTOMÁTICO EN DATOS SIN ETIQUETADO EN PROBLEMAS DE PRIORIZACIÓN GENÉTICA

El tercer caso de uso profundiza en uno de los escenarios más problemáticos y a la vez más habituales para la evaluación de sistemas de IA en biomedicina: aquellos en los que las etiquetas no solo son escasas, sino estructuralmente incompletas. La priorización genética ilustra con claridad cómo los supuestos implícitos de las métricas clásicas fallan cuando la ausencia de evidencia se confunde con evidencia negativa. En estos contextos, evaluar un modelo únicamente en términos de precisión o AUC no solo resulta insuficiente, sino potencialmente engañoso, ya que ignora la incertidumbre inherente al proceso de etiquetado, el coste experimental de los errores y el impacto real de las decisiones de priorización. Este caso pone de relieve que, cuando la IA se utiliza para guiar el descubrimiento científico, las métricas deben capturar no solo la capacidad predictiva del modelo, sino también la calidad del conocimiento generado, la eficiencia computacional y la sostenibilidad del proceso, dimensiones críticas que permanecen invisibles bajo los esquemas de evaluación tradicionales.

La priorización genética constituye una actividad esencial en la investigación biomédica y genómica. Su objetivo es identificar los genes con mayor probabilidad de estar vinculados a un proceso biológico, fenotipo o patología de interés. El empleo de estas metodologías optimiza los esfuerzos del laboratorio, permitiendo que los equipos experimentales se concentren en los candidatos más prometedores, lo cual genera ahorros significativos tanto de tiempo como de recursos económicos y equipamiento. Algunos de los obstáculos con los que se enfrentan los enfoques convencionales de priorización son la escasez y la alta dimensionalidad de los datos genómicos. En este caso de uso, veremos un ejemplo en el que los modelos de aprendizaje automático (AA) han ofrecido una vía prometedora para superar estas dificultades.

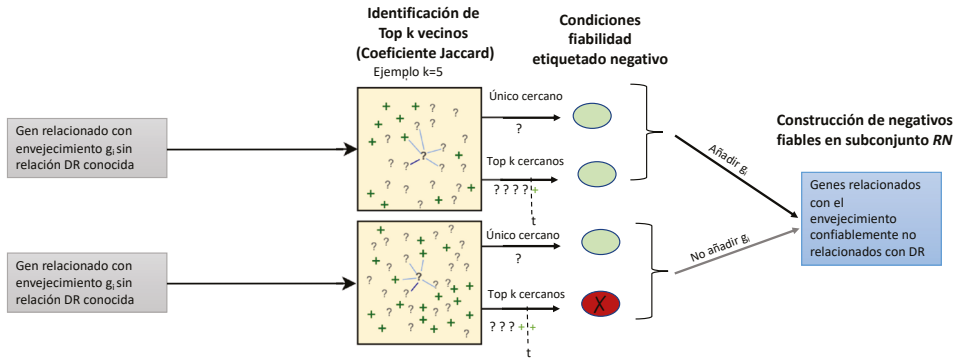
En concreto, nos centraremos en el estudio realizado en (Vega *et al*., 2022), que aplicó AA para hallar genes asociados al envejecimiento que pudieran estar vinculados a la restricción dietética (Dr). La restricción dietética consiste en reducir la ingesta de alimentos sin causar desnutrición y es una de las intervenciones más estudiadas para retrasar

el envejecimiento y mejorar la salud a largo plazo. Se sabe que influye en procesos celulares clave y en numerosos genes, pero identificar exactamente qué genes están implicados es lento y costoso, ya que requiere experimentos de laboratorio complejos. El mencionado estudio utilizó todos los genes sin evidencia conocida de relación con la DR como ejemplos negativos de entrenamiento. Sin embargo, está claro que la carencia de evidencia no equivale a la prueba de la no relación. Esta situación genera un paradigma de datos positivos y sin etiquetar (PU-Positive Unlabeled), donde una parte de los datos tiene una etiqueta positiva confirmada, mientras que el resto (sin etiquetar) se asume como negativo, aunque en realidad algunos podrían ser positivos no detectados. Precisamente, los genes positivos sin etiquetar son el foco de la priorización.

Este trabajo aborda este desafío de datos PU mediante una nueva técnica de aprendizaje PU "basada en distancias" (Paz *et al.*, 2024). El objetivo es perfeccionar los modelos de AA existentes para la priorización genética. El método propuesto, cuyo funcionamiento general se muestra en la figura 8, opera en dos pasos principales:

FIGURA 8

SELECCIÓN DE GENES NEGATIVOS FIABLES BASADA EN SIMILITUD EN EL MÉTODO PROPUESTO



Fuente: Elaboración propia.

- Identificación de negativos confiables: en esta primera etapa, se recurre a un clasificador pseudo-k NN para identificar genes que no están relacionados con la restricción dietética, para ser usados como referencias fiables (ejemplos negativos robustos). Estos son aquellos en los que la mayoría de sus genes vecinos no están relacionados con la DR.

- Entrenamiento del clasificador: Los genes no-DR identificados y los genes DR previamente confirmados se usan para entrenar un ensemble de árboles binarios como modelo de clasificación. De esta forma, el método genera una lista priorizada de nuevos genes candidatos para futuras investigaciones biológicas.

Los resultados son muy positivos: el método no solo mejora la precisión respecto a enfoques anteriores, sino que además reduce el coste computacional y el impacto ambiental, llegando a disminuir las emisiones de CO_2 asociadas al cálculo hasta en un 40 %. Gracias a este enfoque, el estudio identifica varios genes nuevos con un papel potencial en los mecanismos de la restricción dietética, respaldados por evidencias en la literatura científica (tabla 4). Es interesante mencionar que, tras la publicación de nuestro modelo, los genes IRS1 e IRS2 fueron validados como relacionados con la restricción dietética en Drosophila melanogaster.

TABLA 4

GENES CON MAYOR PROBABILIDAD PREDICHA DE RELACIÓN CON LA RESTRICCIÓN DIETÉTICA (DR), COMPARANDO EL MÉTODO EXISTENTE SIN PU (IZQUIERDA) Y EL MÉTODO PROPUESTO CON APRENDIZAJE PU (DERECHA)

Método existente (no-PU)		*Método con aprendizaje PU*	
Gen	**Prob. DR**	**Gen**	**Prob. DR**
GOT2	0.86	*TSC1*	0.97
GOT1	0.85	*GCLM*	0.94
TSC1	0.85	IRS1	0.93
CTH	0.85	PRKAB1	0.92
GCLM	0.82	PRKAB2	0.90
IRS2	0.80	PRKAG1	0.90
SENS2	0.80	*IRS2*	0.90

Este trabajo demuestra cómo una IA mejor diseñada, más consciente de la incertidumbre de los datos, es capaz de mitigar algunas de las limitaciones inherentes a los mismos (naturaleza PU, escasez, alta dimensionalidad) y, por lo tanto, puede convertirse en una aliada clave para la biología del envejecimiento: acelera el descubrimiento científico, reduce costes y hace la investigación más sostenible. Además, logra reducir la complejidad computacional de los métodos previos al mismo tiempo que mantiene su capacidad de explicabilidad.

De forma análoga a los casos de estudio anteriores, este trabajo pone de manifiesto que evaluar sistemas de IA únicamente en términos de precisión o rendimiento predictivo resulta insuficiente. La incorporación de criterios como la calidad del etiquetado, la eficiencia computacional y el impacto medioambiental revela dimensiones del desempeño que permanecen invisibles bajo las métricas tradicionales. En contextos científicos y biomédicos, donde los datos son incompletos, costosos y energéticamente exigentes, se hace imprescindible avanzar hacia nuevas métricas que integren rendimiento, fiabilidad y sostenibilidad, permitiendo comparar modelos no solo por lo que predicen, sino por cómo, con qué coste y con qué impacto lo hacen. Este cambio de perspectiva es clave para una IA verdaderamente responsable y orientada al uso real, especialmente en entornos críticos.

5. MÉTRICA PARA SESGOS EN REGRESIÓN DIÁDICA

El último caso de uso aborda de manera directa uno de los problemas más sutiles y menos visibles de la evaluación en inteligencia artificial: qué ocurre cuando una métrica ampliamente aceptada optimiza el promedio a costa de fallar sistemáticamente en los casos más relevantes.

Definamos primero el contexto en el que trabajaremos: el de los datos diádicos. Los datos diádicos representan información asociada a un par ordenado o no ordenado de entidades (por ejemplo, usuario–usuario, documento–documento, droga–proteína), capturando propiedades de la relación entre ellas más que características aisladas de cada entidad. Ejemplos típicos son las interacciones entre usuarios en redes sociales, las opiniones de usuarios en plataformas como la que vimos anteriormente en TripAdvisor, o las relaciones entre genes y proteínas.

La regresión diádica se refiere a tareas de predicción en las que la variable objetivo está definida sobre pares de entidades y se estima a partir de atributos de cada entidad y/o de su relación. En este tipo de problemas, donde las predicciones se asocian a pares de entidades y tienen consecuencias directas en ámbitos sensibles, las métricas de error global han funcionado durante años como un estándar incuestionado. Sin embargo, este caso muestra que dichas métricas no solo son insuficientes, sino que pueden ocultar comportamientos profundamente problemáticos desde el punto de vista de la equidad, el riesgo y el impacto social. A diferencia de los casos anteriores, aquí no se trata únicamente de qué se mide, sino de qué sesgos estructurales permanecen invisibles cuando no disponemos de la métrica adecuada.

Los modelos de regresión diádica, diseñados para predecir valores reales asociados a pares de entidades, constituyen un componente fundamental de numerosos sistemas de IA actuales. Su aplicación más conocida se encuentra en los sistemas de recomendación (de los que ya hemos visto un caso de uso anteriormente), donde se utilizan para estimar valoraciones o preferencias usuario–ítem, pero su alcance se ha extendido progresivamente a dominios de alto impacto, como la farmacología personalizada, la biomedicina o la modelización de fenómenos económicos y financieros. En estos contextos, las predicciones generadas por los modelos no solo informan decisiones, sino que pueden tener consecuencias directas sobre personas, tratamientos o recursos, lo que hace especialmente relevante analizar cómo se comportan dichos modelos.

A pesar de la complejidad inherente a los datos diádicos —caracterizados por su alta dispersión, distribuciones locales heterogéneas y problemas como el comienzo en frío—, la evaluación de los modelos que operan sobre ellos se ha apoyado históricamente, como ya explicamos anteriormente, en métricas de error global, principalmente RMSE y MAE. Estas métricas, ampliamente consolidadas tras su adopción en hitos como el Netflix Prize (Bennett and Lanning, 2007), ofrecen una visión agregada

del error medio del sistema y han servido como estándar de comparación durante años. Sin embargo, esta perspectiva global oculta aspectos esenciales del comportamiento del modelo cuando se analizan las predicciones a nivel de usuario, ítem o par concreto (Chen, 2017). Los modelos de regresión diádica de última generación tienden a sesgar sus predicciones hacia los valores medios observados de las entidades implicadas, un fenómeno que denominamos sesgo de excentricidad (Paz-Ruza *et al.*, 2025). Este comportamiento permite reducir el error global, pero lo hace a costa de degradar severamente la calidad predictiva en los casos menos frecuentes o más atípicos, que son precisamente los más relevantes en escenarios sensibles desde el punto de vista ético o social. En situaciones extremas, el modelo puede llegar a ofrecer un rendimiento peor que el aleatorio en estos casos críticos, sin que las métricas tradicionales sean capaces de detectarlo.

Nuestros experimentos en ámbitos como la recomendación de contenidos, la predicción de respuesta a fármacos o la estimación de valor económico ponen de manifiesto que RMSE y MAE son insuficientes para revelar estos sesgos estructurales, ya que no distinguen entre errores cometidos en situaciones triviales y errores en predicciones excéntricas con alto impacto potencial. Para abordar esta limitación, proponemos una nueva medida, que llamamos EAUC (*Eccentricity Area Under the Curve*), una métrica diseñada específicamente para capturar cómo varía el error del modelo en función de la excentricidad de los ejemplos, permitiendo así detectar y cuantificar de forma fiable la injusticia algorítmica inducida por el sesgo de excentricidad.

Brevemente, nuestro trabajo, que puede leerse en detalle en (Paz-Ruza *et al.*, 2025), realiza las aportaciones siguientes, que pueden verse resumidas en la figura 9.

- Mostramos que los modelos actuales para regresión diádica sesgan sus predicciones al valor medio observado de las interacciones de cada usuario e ítem, una cuestión que puede ser grave especialmente en entornos críticos. A este fenómeno le llamamos *sesgo de excentricidad*.

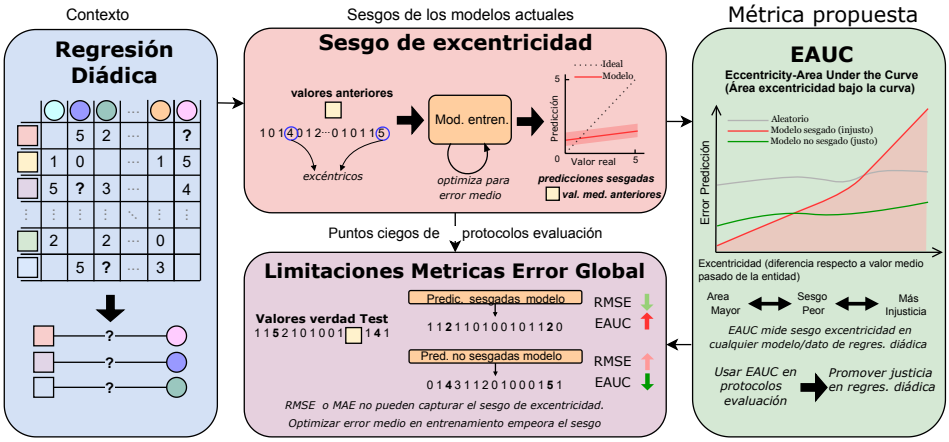
- Demostramos que este sesgo provoca injusticia algorítmica y riesgos en el uso de estos sistemas en tres ámbitos de uso críticos concretos: recomendación de contenidos, farmacología y biomedicina, y mercado y finanzas.

- Demostramos que las medidas de error clásicas, como RMSE o MAE, no detectan estos sesgos en conjuntos de datos reales en los ámbitos anteriores.

- Proponemos una nueva métrica de evaluación, EAUC (*Eccentricity-Area under the Curve*) y un protocolo de evaluación basado en el *sesgo de excentricidad* que permite de forma fiable detectar y cuantificar dicho sesgo en todos los modelos y conjuntos de datos estudiados.

FIGURA 9

RESUMEN DE LAS CONTRIBUCIONES DEL TRABAJO, DE IZQUIERDA A DERECHA: QUÉ ES LA REGRESIÓN DIÁDICA, DEFINICIÓN DEL *SESGO DE EXCENTRICIDAD* EN LA REGRESIÓN DIÁDICA, LAS LIMITACIONES DE LAS MÉTRICAS DE ERROR GLOBALES COMO RMSE O MAE PARA EVALUAR DE MANERA INTEGRAL DICHAS TAREAS, Y LA PROPUESTA DE EAUC COMO UNA MÉTRICA NOVEDOSA QUE PUEDE CUANTIFICAR EL SESGO DE EXCENTRICIDAD (Y LA CONSIGUIENTE INJUSTICIA ALGORÍTMICA) QUE SUFRE UN MODELO DE REGRESIÓN DIÁDICA



Fuente: Elaboración propia.

- Finalmente, proponemos técnicas para prever la tendencia de cualquier conjunto de datos de regresión diádica a inducir este sesgo en modelos entrenados sobre el mismo, así como potenciales aproximaciones para mitigarlo en el futuro.

Este último caso de estudio refuerza una idea central que atraviesa todos los ejemplos analizados en este trabajo: la evaluación de sistemas de IA no puede seguir basándose exclusivamente en métricas de rendimiento promedio. En tareas complejas y de alto impacto, resulta imprescindible avanzar hacia nuevas métricas que incorporen dimensiones como el sesgo, la equidad y el comportamiento del modelo en casos críticos, complementando —y en algunos casos cuestionando— los criterios tradicionales de optimización. Solo mediante este cambio de enfoque será posible desarrollar sistemas de inteligencia artificial realmente fiables, justos y alineados con los requisitos éticos y sociales de sus contextos de aplicación. Este caso de estudio ilustra con especial claridad una conclusión clave de este capítulo: las métricas no son neutrales y elegir mal cómo evaluamos un sistema puede legitimar comportamientos algorítmicos injustos bajo una apariencia de buen rendimiento. La introducción de EAUC demuestra que es posible diseñar métricas que hagan visibles sesgos sistemáticos ignorados por los indicadores clásicos, permitiendo una evaluación más alineada con los riesgos reales de aplicación.

En el contexto del AI Act y de la creciente exigencia de análisis de impacto y equidad, este tipo de métricas no debe entenderse como un complemento opcional, sino como una condición necesaria para una IA responsable. Evaluar no es solo medir el error medio, sino comprender quién falla el sistema, cuándo falla y con qué consecuencias, especialmente en aquellos casos que más importan.

6. CONCLUSIONES

En conjunto, los cuatro casos analizados ponen de manifiesto una idea fundamental: la inteligencia artificial no es neutral y no puede serlo. Los datos que alimentan los modelos reflejan el mundo real, con todas sus desigualdades, asimetrías y sesgos, y los sistemas de IA encarnan necesariamente las decisiones, prioridades e intereses de las organizaciones que los diseñan y despliegan. En este contexto, una evaluación basada casi exclusivamente en métricas de rendimiento técnico —como la precisión, la exactitud o el error medio— resulta claramente insuficiente para comprender el impacto real de estos sistemas sobre las personas, la sociedad y el entorno. Aquello que no medimos no desaparece: la toxicidad en plataformas digitales, los sesgos inducidos por datos incompletos o desbalanceados, la opacidad en la toma de decisiones automatizadas o los costes energéticos y materiales asociados al entrenamiento y uso de modelos de IA continúan existiendo, aunque permanezcan invisibles para las métricas tradicionales. Esta invisibilidad no es inocua: puede traducirse en daños sociales, discriminación, pérdida de confianza o impactos ambientales significativos. Por ello, se hace imprescindible avanzar hacia nuevas métricas que permitan una evaluación integral y responsable de la IA, incorporando dimensiones como la equidad, la explicabilidad, la trazabilidad de las decisiones, la sostenibilidad y la anticipación de riesgos.

La ausencia de estándares claros y, en lo posible, consolidados, en estas dimensiones constituye un desafío adicional que no es únicamente técnico, sino también ético, social y político. Diseñar y desplegar sistemas de IA justos y sostenibles requiere, por tanto, no solo innovación metodológica y tecnológica, sino un compromiso deliberado y coordinado de toda la comunidad: investigadores, reguladores y sociedad deben colaborar para que la IA deje de ser una "caja negra" optimizada únicamente para el rendimiento, y se convierta en una tecnología alineada con valores humanos y orientada al bien común.

Referencias

Bennett, J., and Lanning, S. (2007). The Netflix prize. In Proc. KDD Cup Workshop, 35.

Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., and Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599, 128096. https://doi.org/10.1016/j.neucom.2024.128096.

Chen, M. (2017). Performance evaluation of recommender systems. *Int. J. Performability Eng.*, *vol. 13*, no. 8, 1246.

Deters, H., Droste, J., Obaidi, M., and Schneider, K. (2025). Exploring the means to measure explainability: Metrics, heuristics and questionnaires. *Information and Software Technology*, 181, 107682, https://doi.org/10.1016/j.infsof.2025.107682.

Diez, J., Perez-Nunez, P., Luaces, O., Remeseiro, B., and Bahamonde, A. (2020). Towards explainable personalized recommendations by learning from users' photos. *Inform. Sci.*, 520, 416-430.

Glickman, M., Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional, and social judgements. *Nat Hum Behav,* 9, 345–359. https://doi.org/10.1038/s41562-024-02077-2.

Goellner, S., Tropmann-Frick, M., and Brumen, B. (2024). Responsible Artificial Intelligence: A Structured Literature Review. ArXiv, abs/2403.06910.

Hanu, L. and [Unitary team]. (2020). Detoxify. Howpulished= github. https://github.com/unitaryai/detoxify

Imani, M., Joudaki, M., Bagheri, A., and Arabnia, H. R. (2026). Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers. *Technologies*, 14, 54. https://doi.org/10.3390/technologies14010054

Palumbo, G., Carneiro, D., and Alves, V. (2025). Objective metrics for ethical AI: a systematic literature review. *Int J. Data Sci. Anal.,* 20, 247–267. https://doi.org/10.1007/s41060-024-00541-w.

Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Cancela, B., and Eiras-Franco, C. (2024). Sustainable transparency on recommender systems: Bayesian ranking of images for explainability. *Information Fusion*, Volume 111, 102497. https://doi.org/10.1016/j.inffus.2024.102497.

Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Cancela, B., and Eiras-Franco, C. (2025). Beyond RMSE and MAE: Introducing EAUC to Unmask Hidden Bias and Unfairness in Dyadic Regression Models. *IEEE Transactions on Neural Networks and Learning Systems*, *vol. 36*, no. 11, 19619-19630, Nov. doi: 10.1109/TNNLS.2025.3593059.

Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., and Eiras-Franco, C. (2025). Predictively combating toxicity in Health-related online discussions through Machine Learning. Proc. International Joint Conference on Neural Networks, IJCNN 2025. DOI: 10.1109/IJCNN64981.2025.11228479.

Paz-Ruza, J., Eiras-Franco, C., Guijarro-Berdiñas, B., and Alonso-Betanzos, A. (2022). Sustainable Personalisation and Explainability in Dyadic Data Systems. *Procedia Computer Science*, Volume 207, 1017-1026, 2022. https://doi.org/10.1016/j.procs.2022.09.157.

Paz-Ruza,J., Freitas, A. A., Alonso-Betanzos, A., and Guijarro-Berdiñas, B. (2024). Positive-Unlabelled learning for identifying new candidate Dietary Restriction-related genes among ageing-related genes. *Computers in Biology and Medicine*, 180, 108999. https://doi.org/10.1016/j.compbiomed.2024.108999.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency , FAT* '20, (New York, NY, USA)*, 33–44, Association for Computing Machinery.

Schilke, O., and Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, 104405. https://doi.org/10.1016/j.obhdp.2025.104405.

Schwartz, R., Dodge, J., Smith, N. A., Etzioni, O. (2020). Green AI. *Commun. ACM, 63*(12), 54–63.

Vega, M., Gustavo, D., Bespalov, V., Zheng, Y., Freitas, A., and De Magalhaes, J. P. (2022). Machine learning-based predictions of dietary restriction associations across ageing-related genes. *BMC bioinformatics*, 23, 1–28.

CAPÍTULO II

Aprendizaje federado para una sociedad de máquinas

Senén Barro*
José Miguel Burés
Roberto Iglesias

La inteligencia artificial a menudo se concibe como el proyecto de diseñar agentes artificiales capaces de tomar decisiones de forma autónoma y actuar de forma eficiente en entornos cada vez más complejos. Cuanto más sofisticadas se hagan las tareas que los humanos deleguemos en dichos agentes, menos podremos supervisar sus decisiones y mayor será el impacto (positivo o negativo) que tendrán sus acciones en nuestras vidas. Hay una concienciación creciente de que la autonomía de los sistemas IA ha de desempeñarse de forma responsable y, en particular, que sus decisiones y acciones han de estar alineadas con los valores (éticos) de la sociedad. En este capítulo se van a repasar algunas técnicas para gobernar los sistemas multiagente, y presentar modelos recientes para que los agentes IA puedan desempeñar su autonomía de forma responsable, alineando sus acciones con los valores éticos de la sociedad.

Palabras clave: aprendizaje automático, aprendizaje federado, privacidad de los datos, datos no IID, deriva de concepto, personalización del aprendizaje.

* Este trabajo ha recibido apoyo financiero de la Xunta de Galicia — Consellería de Educación, Ciencia, Universidades e Formación Profesional (acreditación de Centro de investigación de Galicia 2024-2027 ED431G-2023/04 y acreditación de Grupo de Referencia Competitiva (2022-2025, Project ED431C 2022/19 CONSOLIDACIÓN 2022 GRC-GSI) y de la Unión Europea (Fondo Europeo de Desarrollo Regional — FEDER).

1. DEL RAZONAMIENTO AUTOMÁTICO AL APRENDIZAJE AUTOMÁTICO

A lo largo de su historia, la inteligencia artificial ha seguido dos grandes aproximaciones para abordar la resolución de problemas. La primera se apoya en la representación explícita del conocimiento y el razonamiento automático; la segunda, en el aprendizaje a partir de datos mediante métodos estadísticos y de optimización, fundamentalmente. Aunque a menudo se presentan como paradigmas contrapuestos, ambas aproximaciones han coexistido desde los orígenes del campo. Lo que ha cambiado con el tiempo es cuál de ellas ha ocupado una posición predominante en la investigación y en las aplicaciones prácticas.

En sus inicios, durante las décadas de 1950 y 1960, la IA se desarrolló fundamentalmente bajo una perspectiva simbólica. La hipótesis central era que el comportamiento inteligente podía entenderse como la manipulación de símbolos de acuerdo con reglas formales. Tuvieron una especial atención los problemas asociados a espacios de estados en los que la resolución de un problema significa encontrar un camino que nos lleve desde el estado de partida a otro estado, normalmente no único, que podamos considerar como una buena solución, si no la óptima, al problema. Los juegos de mesa son un ejemplo paradigmático, y dentro de estos, el ajedrez. En ese proceso de encontrar un estado solución, normalmente es inviable explorar computacionalmente todas las opciones, así que se hizo común el uso de heurísticas diseñadas por humanos para guiar la búsqueda de una solución. En este contexto se consolidaron conceptos como la resolución de problemas mediante búsqueda, la demostración automática de teoremas y el uso de la lógica como lenguaje de representación del conocimiento. Los trabajos fundacionales de Newell, Simon y McCarthy establecieron una visión de la IA como una disciplina orientada a modelar el razonamiento humano a través de estructuras simbólicas y procesos inferenciales (McCarthy, 1959; Newell y Simon, 1976).

Con el paso del tiempo, se hizo evidente que estos enfoques funcionaban bien en dominios pequeños y altamente estructurados, pero tenían dificultades para escalar a problemas del mundo real, donde el conocimiento relevante es ingente, incompleto y a menudo incierto. Esta constatación condujo, en los años setenta y principios de los ochenta, al auge de los sistemas expertos (a veces denominados también sistemas basados en conocimiento). En lugar de aspirar a un razonamiento general, estos sistemas se centraban en capturar el conocimiento específico de un dominio mediante formas computacionalmente tratables. Dicho conocimiento podía proceder de fuentes escritas, como libros o publicaciones científicas, o directamente de expertos humanos. La promesa era pragmática: si se lograba formalizar lo que saben los especialistas sobre un tema, el sistema podría replicar —al menos parcialmente— su capacidad de resolución de problemas en su ámbito de competencia. Durante este periodo, la IA simbólica alcanzó su mayor impacto industrial, con aplicaciones en diagnóstico médico, configuración de sistemas y soporte a la decisión (Shortliffe, 1976).

Sin embargo, este enfoque puso de manifiesto limitaciones estructurales importantes. La adquisición y el mantenimiento del conocimiento resultaron ser procesos costosos y complejos, y los sistemas tendían a comportarse de forma inadecuada ante situaciones no previstas en el conocimiento disponible, a menudo muy incompleto. Además, el tratamiento de la incertidumbre reveló las carencias de una lógica puramente determinista, lo que motivó la incorporación de modelos probabilísticos y métodos de razonamiento bajo incertidumbre, como las redes bayesianas (Pearl, 1988). Estas dificultades, combinadas con expectativas sobredimensionadas —algo común durante el desarrollo de la IA— desembocaron en periodos de desilusión y reducción de financiación, conocidos como "inviernos de la IA".

A finales de los años ochenta y durante la década de 1990 se produjo un cambio gradual pero profundo en la orientación del campo. El aumento de la capacidad de cómputo y la creciente disponibilidad de datos digitales favorecieron el desarrollo de métodos basados en el *aprendizaje automático*. En lugar de codificar explícitamente el conocimiento, estos enfoques se centraban en aprender modelos a partir de ejemplos, evaluando su calidad mediante criterios empíricos como el error de generalización. Se consolidó así una visión más cercana a la estadística y al reconocimiento de patrones, con técnicas como los árboles de decisión, los modelos probabilísticos y las máquinas de vectores soporte (Quinlan, 1993; Vapnik, 1995).

Las redes neuronales, presentes desde los orígenes de la disciplina, experimentaron en este periodo un nuevo impulso, especialmente en tareas de percepción como el reconocimiento de caracteres. Aun así, durante muchos años su uso estuvo limitado por dificultades de entrenamiento y por la falta de datos y cómputo suficientes (LeCun *et al.,* 1998). No fue hasta mediados de la década de 2000 cuando una serie de avances técnicos permitió entrenar modelos más grandes y complejos, sentando las bases de lo que posteriormente se conocería como aprendizaje profundo (Hinton *et al.,* 2006).

El punto de inflexión que marca el predominio actual del enfoque basado en datos suele situarse en torno a 2012, con resultados espectaculares en visión por computador obtenidos mediante redes neuronales profundas entrenadas a gran escala (Krizhevsky *et al*., 2012). Desde entonces, el aprendizaje profundo ha pasado a dominar áreas como la visión, el procesamiento del lenguaje natural y, en combinación con el aprendizaje por refuerzo, ciertos problemas de control y toma de decisiones. Sistemas capaces de aprender directamente a partir de percepciones de bajo nivel y de optimizar su comportamiento mediante la interacción con el entorno ilustran la potencia de esta aproximación (Mnih *et al., 2015;* Silver *et al.*, 2016).

No obstante, este predominio no implica la desaparición del enfoque simbólico. Por el contrario, las limitaciones de los sistemas puramente basados en datos —dependencia de grandes volúmenes de información, dificultades de interpretación, falta de garantías formales— han reavivado el interés por integrar conocimiento explícito, razonamiento y

aprendizaje. En la práctica, muchas de las arquitecturas más avanzadas combinan componentes aprendidos con mecanismos clásicos de búsqueda, planificación o restricción lógica, dando lugar a enfoques híbridos que tratan de aprovechar lo mejor de ambos paradigmas.

Desde una perspectiva histórica, puede afirmarse que la IA ha oscilado entre periodos de predominio simbólico y periodos dominados por el aprendizaje a partir de datos, sin que ninguno de los dos enfoques haya sido nunca completamente excluyente. La situación actual refleja más bien una convergencia: el reconocimiento de que la inteligencia artificial robusta y confiable probablemente requiera tanto de la capacidad de aprender de la experiencia como la de razonar con estructuras de conocimiento explícitas, dependiendo del problema y del contexto de aplicación.

2. ¿QUÉ ES EL APRENDIZAJE AUTOMÁTICO Y CUÁL ES SU ORIGEN?

El ámbito del aprendizaje automático está integrado en el de la inteligencia artificial, y se orienta al desarrollo de algoritmos que mejoran automáticamente a través de la experiencia. Me parece especialmente elegante la definición dada por Tom Mitchell, brillante investigador de la IA: "Un sistema aprende de la *experiencia* X con respecto a una *tarea* T y con una medida de *rendimiento* P, si su rendimiento para la tarea T, medido mediante P, mejora con la experiencia X" (Mitchell, 1997).

Abunda la idea equivocada de que el ámbito del aprendizaje automático es bastante reciente, pero no es así. Algunas de las bases matemáticas están establecidas desde hace un par de siglos. Sirva de ejemplo el ajuste de puntos a una recta por mínimos cuadrados (regresión lineal). La primera forma de regresión lineal documentada fue el método de los mínimos cuadrados publicado por Legendre en 1805. Es cierto que ahí están las bases matemáticas e implícito un algoritmo que aprende una función que permite predecir el valor de una variable independiente dado el valor de una o más variables dependientes, pero no se trata de una implementación en una máquina que automatice el proceso de aprender dicha función.

En todo caso, hace bastantes décadas que el aprendizaje automático como tal dio sus primeros pasos. El primer programa informático que empleó aprendizaje automático fue el juego de damas desarrollado por Arthur Samuel en 1952 (Samuel, 1959). Este programa, ejecutado en una computadora IBM 701, se considera pionero porque fue capaz de mejorar su desempeño a medida que jugaba más y más partidas.

Samuel diseñó su *software* para que ajustara sus estrategias en función de las partidas previas, afinando su capacidad de juego sin intervención humana directa. El programa no solo calculaba movimientos óptimos con reglas predefinidas, sino que también

empleaba técnicas de refinamiento basadas en la experiencia, un concepto fundamental en el aprendizaje automático (recordemos la definición de aprendizaje dada por Mitchell). Aunque rudimentario, su enfoque inspiró futuros algoritmos de aprendizaje por refuerzo, esenciales en la inteligencia artificial moderna, en particular en los modelos grandes de lenguaje o LLM.

El trabajo de Samuel sentó las bases del aprendizaje automático como disciplina científica, demostrando que los programas podían "aprender de la experiencia" en lugar de seguir reglas estrictamente programadas.

3. APRENDIZAJE SUPERVISADO, NO SUPERVISADO Y POR REFUERZO

El interés por el aprendizaje automático es tal que la mayoría de los investigadores en IA están implicados de un modo u otro en él. No es de extrañar, así que la mayor parte de las publicaciones científicas en los congresos de IA se sitúen en la órbita del aprendizaje automático (y dentro de este, de un modo notorio, en los modelos grandes de lenguaje, o LLM). Es cierto que en la mayor parte de los casos no se trata de propuestas realmente originales, sino de variantes, muchas veces de escasísimo valor, o ninguno, de modelos base en aprendizaje automático, como pueden ser los clasificadores (Fernández-Delgado *et al.*, 2014) o regresores (Fernández-Delgado *et al.*, 2019).

En general, casi todos los algoritmos y aplicaciones de aprendizaje automático se sitúan en tres categorías principales, dependiendo de cómo se guía el proceso de aprendizaje: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. Esta distinción es clásica, pero sigue siendo conceptualmente útil porque refleja tres maneras distintas de relacionar datos, objetivos y funciones a optimizar —aprendizaje propiamente dicho—, lo que ayuda a entender qué tipo de problemas puede abordar cada enfoque y cuáles son sus límites.

El *aprendizaje supervisado* es, históricamente y aún hoy, la forma más extendida y mejor comprendida. En este paradigma, el sistema dispone de un conjunto de ejemplos en los que cada entrada está asociada a una salida deseada, decidida de antemano y en coherencia con el problema a resolver. El aprendizaje se guía explícitamente por esa señal externa: el modelo ajusta sus parámetros para minimizar la discrepancia entre sus predicciones y la realidad buscada —etiquetas con la respuesta deseada para los datos de entrada—. Problemas de clasificación, regresión y reconocimiento de patrones encajan naturalmente en este marco, y gran parte del éxito reciente del aprendizaje profundo se apoya en formulaciones supervisadas a gran escala. Su fortaleza principal es la claridad del objetivo: se detalla lo que se espera del sistema. Su principal limitación es también evidente: requiere grandes cantidades de datos etiquetados, cuya obtención suele ser costosa, lenta o directamente inviable.

Un ejemplo claro y simple de problema resoluble con aprendizaje supervisado es la clasificación de correo electrónico (spam/no spam), en el que partiríamos de un conjunto de entrenamiento formado por cientos o miles de ejemplos de correos etiquetados, según el caso, como deseados o no deseados. El modelo aprende a generalizar a partir de estos ejemplos para clasificar correos nuevos, ya en condiciones reales de operación.

En contraste, el *aprendizaje no supervisado* prescinde de etiquetas explícitas. El sistema recibe únicamente los datos de entrada y debe descubrir por sí mismo regularidades, estructuras o representaciones internas que capturen aspectos relevantes de esos datos, útiles para abordar el problema en cuestión. Aquí, el proceso de aprendizaje no está guiado por una noción externa de "respuesta correcta", sino por criterios internos como la compacidad, la reconstrucción, la independencia o la probabilidad de los datos observados. Históricamente, este tipo de aprendizaje ha estado ligado a tareas como el agrupamiento, la reducción de dimensionalidad o el modelado de distribuciones. Conceptualmente, su importancia va más allá de estas aplicaciones concretas: el aprendizaje no supervisado encarna la idea de que un sistema puede organizar un micromundo —a través de los datos que lo representan— sin supervisión directa, lo que lo aproxima a ciertos mecanismos de aprendizaje humano y animal. Sin embargo, precisamente por carecer de una señal externa clara, evaluar y controlar lo que se aprende resulta más difícil, y los resultados pueden ser más dependientes de supuestos implícitos del modelo.

Un ejemplo de aplicación diseñable a través del aprendizaje no supervisado es la segmentación de clientes a partir de datos derivados de su comportamiento (compras, frecuencia, gasto). El sistema agrupa clientes con patrones similares sin que nadie defina previamente qué es un "tipo" de cliente. El resultado es una estructura útil para análisis o *marketing*, aunque no impuesta desde fuera. Será tras la identificación de los agrupamientos que se correspondan con distintos perfiles de clientes, cuando estos se asocien con una etiqueta o acción de utilidad en el dominio.

El *aprendizaje por refuerzo* (Sutton y Barto, 2018) introduce una tercera forma de aprendizaje. En este caso, el sistema aprende a partir de su interacción con un entorno, recibiendo señales de recompensa o castigo que evalúan las consecuencias positivas o negativas de sus acciones. A diferencia del aprendizaje supervisado, no se le indica qué acción es correcta en cada situación; solo se le proporciona una valoración escalar de su comportamiento, a menudo con un cierto retraso temporal. El proceso de aprendizaje se guía así por un objetivo global (maximizar la recompensa acumulada) y no por ejemplos individuales etiquetados. Este paradigma es especialmente adecuado para problemas de control, toma de decisiones secuenciales y planificación bajo incertidumbre. Su complejidad conceptual es mayor, ya que el agente debe enfrentarse simultáneamente a la exploración del entorno y a la explotación de lo ya aprendido, y los datos dependen de las propias decisiones del sistema. Hace décadas que en nuestro grupo de investigación usamos este tipo de aprendizaje para que los robots autónomos aprendan a moverse en entornos complejos y dinámicos y a resolver tareas en ellos.

Aunque estas tres formas suelen presentarse como categorías separadas, en la práctica sus fronteras son cada vez más permeables. Existen enfoques *semisupervisados*, que combinan pequeños conjuntos de datos etiquetados con grandes volúmenes de datos sin etiquetar; métodos *autosupervisados*[1], en los que la señal de aprendizaje se construye a partir de la propia estructura de los datos; y sistemas de aprendizaje por refuerzo que incorporan modelos supervisados o representaciones aprendidas de forma no supervisada. Esta hibridación refleja una comprensión más madura del problema: guiar el aprendizaje no es una decisión binaria, sino un diseño continuo que depende del tipo de señales o datos disponibles, del coste de obtenerlos y del nivel de autonomía deseado para el sistema.

Desde una perspectiva conceptual, estas tres formas de aprendizaje corresponden a tres respuestas distintas a una misma pregunta fundamental: ¿de dónde proviene la información que orienta la mejora del modelo? En el aprendizaje supervisado, procede de un maestro explícito; en el no supervisado, de la propia estructura de los datos; y en el aprendizaje por refuerzo, de la interacción con el entorno —físico o digital— y sus consecuencias. Entender esta distinción es clave para situar cualquier técnica concreta de aprendizaje automático dentro de un marco más amplio y para elegir, de manera informada, el paradigma adecuado para cada problema.

Seguramente el lector, aunque no sea un especialista en inteligencia artificial, habrá oído hablar del *aprendizaje profundo* y de su extraordinaria influencia en el *boom* actual de la IA. Es importante aclarar que no se trata de un nuevo tipo de aprendizaje en el mismo sentido que el aprendizaje supervisado, no supervisado o por refuerzo. Como hemos dicho, estas categorías describen cómo se guía el proceso de aprendizaje a través de distintas tipologías de datos (etiquetados, sin etiquetar y asociados a recompensas o penalizaciones, según el resultado de las decisiones tomadas por el sistema). En cambio, el aprendizaje profundo (LeCun *et al.*, 2015) describe una familia de modelos y representaciones, no el tipo de datos disponibles durante el aprendizaje. Se refiere al uso de redes neuronales con muchas capas que aprenden representaciones jerárquicas mediante optimización por gradiente (Rumelhart *et al.*, 1986). Desde este punto de vista, el aprendizaje profundo es una opción arquitectónica y algorítmica, no un paradigma de aprendizaje en sí mismo. De hecho, en el aprendizaje profundo pueden combinarse los distintos tipos de aprendizaje antes descritos.

[1] En el caso de los LLM, su fase inicial de entrenamiento se realiza mediante aprendizaje autosupervisado. En concreto, partiendo de grandes colecciones de texto sin etiquetar, se generan tareas de predicción interna, como predecir la siguiente palabra (o token) a partir del contexto previo, o bien reconstruir partes del texto ocultas. El texto original proporciona simultáneamente la entrada y la "señal de supervisión": no se añade información externa, pero el aprendizaje sí está guiado por un objetivo explícito.

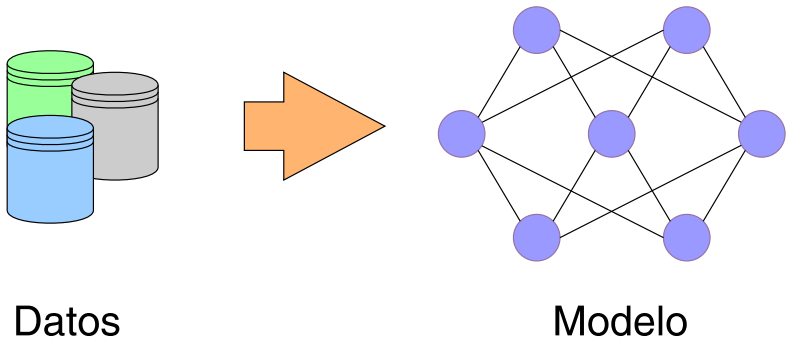
4. DEL APRENDIZAJE LOCAL AL FEDERADO

Pensemos ahora en las posibilidades de aplicación del aprendizaje automático en función de en dónde residen los datos y cómo se accede a ellos durante el entrenamiento. En este sentido, es posible trazar una tipología del aprendizaje automático que resulta especialmente útil para entender tanto las arquitecturas técnicas como las implicaciones prácticas (privacidad, escalabilidad, control, costes). Este eje es relativamente reciente en la historia del campo, pero hoy es central en muchos sistemas reales. En la forma más clásica de aprendizaje automático, los datos se encuentran centralizados en un único dispositivo o repositorio y el entrenamiento se realiza directamente sobre ese conjunto de entrenamiento. Este esquema, que durante décadas fue prácticamente el único considerado, presupone que los datos pueden recopilarse, almacenarse y procesarse en un único entorno computacional. Desde el punto de vista algorítmico, es el escenario ideal, por ser el más simple y el que en principio permite obtener mejores resultados: el modelo se entrena optimizando una función objetivo definida sobre todos los datos disponibles. Gran parte del aprendizaje supervisado, no supervisado y profundo se ha formulado históricamente bajo esta hipótesis. Su principal ventaja es la eficiencia estadística y computacional; su limitación, cada vez más relevante, es que no siempre es viable concentrar datos por razones de volumen, latencia, propiedad o privacidad.

El aprendizaje que puede tener acceso al conjunto entero de datos de entrenamiento, y también de validación, puede ser local, si hay una única fuente o repositorio de datos, o puede ser centralizado si estas son múltiples pero los datos en última instancia están disponibles sin limitaciones para que un sistema central haga su trabajo.

FIGURA 1

ESQUEMA DE APRENDIZAJE LOCAL. SE GENERA UN MODELO ESPECÍFICO PARA PROCESAR LOS DATOS LOCALES



Fuente: Elaboración propia.

Formalizando mínimamente el funcionamiento de un modelo de *aprendizaje local,* podemos partir de una entidad E, que opera en un mundo W, en el que realiza una tarea T, a través de un modelo M, aprendido a partir de su experiencia X, con una medida de rendimiento P:

$$E(X(t)) \rightarrow M(t) \quad [1]$$

El modelo para realizar T puede ser dinámico, buscando mejorar el rendimiento, $P(t)$, a través de nuevas experiencias, $X(t)$, en cuyo caso:

$$E(M(t), X(t)) \rightarrow M(t+\Delta), \text{ con } P(t+\Delta) \geq P(t) \quad [2]$$

Cuando los datos están distribuidos entre múltiples fuentes, pero pueden agregarse o compartirse, al menos parcialmente, aparecen esquemas de aprendizaje centralizado o distribuido.

El aprendizaje centralizado se asume que es posible concentrar los datos. Si es posible juntarlos todos para poder aprender sobre ellos sin más, casi siempre va a ser la mejor opción, salvo si hacerlo de ese modo añade costes muy significativos de comunicación, memoria o demora en la obtención de una primera aproximación a la resolución del problema, pongamos por caso.

En el *aprendizaje centralizado* un conjunto de entidades, $ES = \{E_1, E_2 \cdots E_S\}$, comparte sus experiencias $XS = \{X_1, X_2 \cdots X_S\}$, de modo que es posible calcular directamente un modelo global centralizado, MC, de resolución de la tarea T, a partir de las mismas:

$$EC(\{X_s(t), s = 1, \cdots, S\}) \rightarrow MC(t), \quad [3]$$

tal que, idealmente $PC(t) \geq P_s(t), \forall s = 1, \cdots, S$.

El aprendizaje centralizado para resolver una tarea se calcula generalmente en una arquitectura cliente-servidor y suele seguir el siguiente procedimiento (ver figura 2):

Paso 1. Los clientes envían sus datos al servidor (que se gestionan específicamente para añadir privacidad y seguridad, si es necesario).

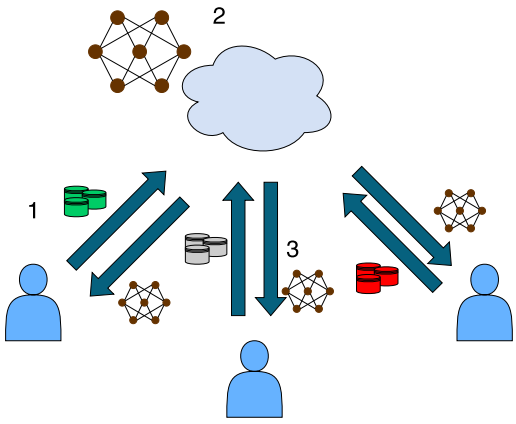
Paso 2. El servidor utiliza todos los datos para aprender una solución para resolver la tarea.

Paso 3. El servidor envía la solución a los clientes (la solución puede adaptarse o no a cada uno de ellos según sus especificidades).

Paso 4. Los clientes adoptan la solución proporcionada por el servidor y operan con ella.

FIGURA 2

APRENDIZAJE CENTRALIZADO, MEDIANTE UNA ARQUITECTURA CLIENTE-SERVIDOR: (1) LOS CLIENTES ENVÍAN SUS DATOS AL SERVIDOR; (2) ESTE CREA UNA SOLUCIÓN GLOBAL CON TODA LA INFORMACIÓN Y (3) LA DISTRIBUYE DE VUELTA A LOS CLIENTES



Fuente: Elaboración propia.

En el *aprendizaje distribuido* el entrenamiento se reparte entre varios nodos de cómputo, que procesan subconjuntos de los datos y sincronizan periódicamente los parámetros del modelo. Esta distribución puede responder a razones puramente técnicas (acelerar el entrenamiento de modelos grandes) o a la naturaleza misma del sistema (datos generados en múltiples localizaciones). Aunque conceptualmente sigue tratándose de un aprendizaje "global", como el centralizado, en este caso no existe un único punto donde residen todos los datos durante el proceso de construcción del modelo solución. Este enfoque se ha vuelto esencial en el entrenamiento de modelos a gran escala, pero no resuelve por sí mismo todas las restricciones que puedan existir en ciertos problemas, singularmente los de privacidad, ya que los datos o sus derivados siguen siendo visibles para la infraestructura central.

Un paso muy significativo en este sentido, y el punto al que queríamos llegar por ser el tema central de este texto, lo constituye el *aprendizaje federado* (McMahan *et al.*, 2017), en el que los datos permanecen en los dispositivos o lugares de generación de los mismos, de modo que en el aprendizaje sobre ellos solo se intercambia información de lo aprendido (funciones o modelos resultantes durante el aprendizaje) en ningún

caso los datos de entrenamiento. Este paradigma responde a restricciones cada vez más frecuentes en contextos como dispositivos personales o aplicaciones en el ámbito de la salud o las finanzas, donde la privacidad de los datos es una cuestión legal, o al menos ética.

Dado un conjunto de entidades $ES = \{E_1, E_2 \cdots E_S\}$ que resuelven la tarea T mediante modelos locales $MS = \{M_1, M_2 \cdots M_S\}$, obtenidos a partir de sus experiencias particulares $MS = \{M_1, M_2 \cdots M_S\}$, nos planteamos obtener un modelo global, MF, mediante una estrategia de aprendizaje federado, que mejore o ayude a mejorar a cualquier modelo local:

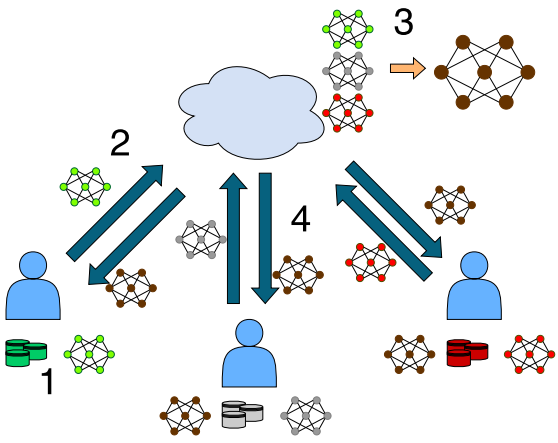
$$EF(\{M_s(t), s = 1, \cdots, S\}) \rightarrow MF(t), \qquad [4]$$

tal que, idealmente, $PF(t) \geq P_s(t), \forall s = 1, \cdots, S$, si t es suficientemente grande.

Las arquitecturas cliente-servidor son especialmente adecuadas para el aprendizaje federado que funciona con conjuntos de datos que comparten el mismo conjunto de variables o características; por ejemplo, los que provienen de un conjunto de estaciones meteorológicas distribuidas por un territorio o de una misma aplicación de móviles instalada en muchos de estos dispositivos. A continuación se describe el procedimiento general para una arquitectura cliente-servidor federada (ver figura 3):

FIGURA 3

ARQUITECTURA CLIENTE-SERVIDOR FEDERADA: 1) ENTRENAMIENTO LOCAL EN CADA CLIENTE, (2) ENVÍO DE MODELOS AL SERVIDOR, (3) COMBINACIÓN PARA CREAR UN MODELO GLOBAL Y (4) DISTRIBUCIÓN DE LA SOLUCIÓN A LOS CLIENTES



Fuente: Elaboración propia.

Paso 1. N clientes aprenden a resolver una tarea en sus conjuntos de datos (normalmente a partir de un modelo inicial proporcionado por el servidor y que suele ser el mismo para todos los clientes, al menos cuando todos ellos tienen que resolver la misma tarea).

Paso 2. Los modelos locales resultantes del aprendizaje local, $\{M_n, n = 1, \cdots, N\}$, se envían al servidor (después de ser tratados específicamente para añadir más privacidad y seguridad, si es necesario).

Paso 3. El servidor utiliza todos los modelos locales para aprender un modelo mejor para resolver la tarea.

Paso 4. El servidor envía la solución a los N clientes (la solución puede adaptarse o no a cada uno de ellos en función de sus especificidades y/o de las tareas de las que son responsables).

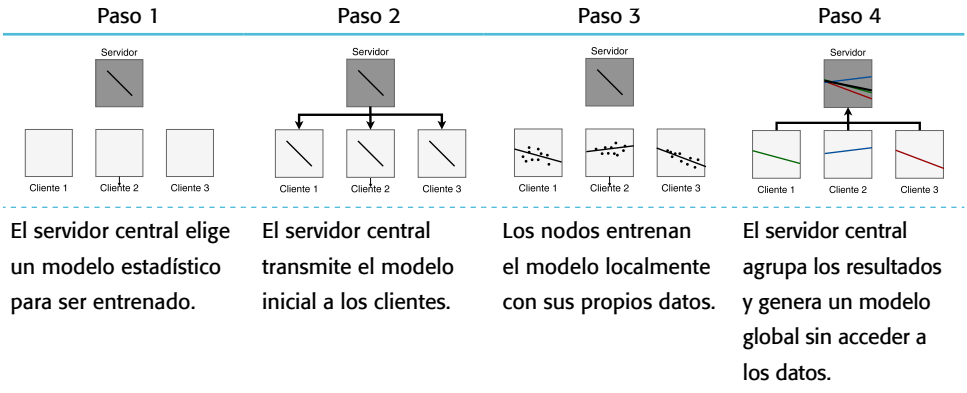
Paso 5. Los clientes asumen la solución proporcionada por el servidor y operan con ella (la solución puede ser común o no a todos ellos).

Normalmente el procedimiento indicado se repite, de modo que los clientes continúan mejorando sus modelos para resolver las tareas de las que son responsables, hasta que se verifica un criterio de parada, como haber alcanzado un rendimiento local y/o colectivo determinado.

Por si no le ha quedado claro cuál es la base del aprendizaje federado, le pondremos un ejemplo simple que creemos que le ayudará a entenderlo mejor. Gráficamente se describe en la figura 4. Imagínese que tiene tres fuentes de datos y quiere aprender la función lineal que los ajusta en cada caso a una recta. En este caso los datos no están disponibles de partida, sino que se trata de un sistema dinámico en el que van entrando continuamente nuevos datos y la función aprendida ha de irse ajustando a ellos a lo largo del tiempo. Además, los datos locales no pueden compartirse por una cuestión de privacidad. En primer lugar, para poder operar integrando las funciones aprendidas y no los datos locales, la función a aprender tendrá que ser única (paso 1). A partir de ahí, todos los aprendedores —clientes— adoptarán esa función como punto de partida previo al ajuste de la misma a los datos locales de los que dispone cada aprendedor (paso 2). Los aprendedores ya operarán con sus respectivas funciones locales (paso 3), que en este caso serán progresivamente ajustadas a los nuevos datos locales. Tras un cierto tiempo de reajuste local continuo de las funciones, los aprendedores envían sus

FIGURA 4

ILUSTRACIÓN DE LA IDEA BASE DEL APRENDIZAJE FEDERADO, EN LA QUE LOS CLIENTES LOCALES AJUSTAN CON SUS DATOS UNA FUNCIÓN OBJETIVO, LA COMPARTEN PARA OBTENER UNA FUNCIÓN DE SÍNTESIS GLOBAL, QUE DE NUEVO UTILIZAN Y AJUSTAN A SUS DATOS LOS APRENDEDORES LOCALES, SIGUIENDO UN PROCESO ITERATIVO



Fuente: Elaboración propia.

respectivas funciones al servidor y este hará una composición de las mismas para obtener una función promedio (paso 4).

Fijémonos que en este ejemplo se asume que los datos en cada sistema local no están dados de antemano como un conjunto único y estático, sino que llegan secuencialmente, y el sistema debe actualizar su modelo de manera incremental. En muchos escenarios reales —sistemas de recomendación, detección de fraude, control— los datos se generan continuamente y no pueden almacenarse o reprocesarse en bloque. En estos casos, el aprendizaje está estrechamente ligado al flujo de datos y a las decisiones sobre qué información conservar, descartar o resumir. Desde esta perspectiva, el "lugar" de los datos es efímero: existen en el momento en que son observados y, tras su uso, pueden desaparecer. Después volveremos a tratar este tema del aprendizaje continuo, tan importante para un sinfín de aplicaciones.

4.1. ¿Qué ocurre cuando no se quiere o no se pueden compartir los datos?

Tenemos muy cerca en el tiempo la pandemia. Seguramente recordará el lector las aplicaciones de seguimiento de contagios (*contact tracing apps*). En general, hubo dos métodos con arquitecturas de gestión de datos diferentes, centralizadas y no centralizadas (Troncoso *et al.,* 2022), en función de garantizar, respectivamente, una mayor utilidad en salud pública o una mayor privacidad de los datos del usuario. España asumió una aplicación denominada Radar COVID, basada en el modelo descentralizado de notificación de exposiciones, alineado con las recomendaciones europeas de privacidad

y con el enfoque adoptado por la mayoría de los países de la Unión Europea (UE). Es cierto que su escasa utilización la hizo inútil en la práctica.

En las arquitecturas no centralizadas, los móviles próximos entre sí intercambiaban identificadores anónimos. Cada móvil disponía así de la huella dejada por aquellos otros móviles que en un momento dado estuvieron cercanos a aquel. Cada persona que enfermase de COVID-19 y dispusiese de esta aplicación en su móvil, debería notificarlo a un servidor central, que se encargaría de ir registrando los identificadores generados por los móviles de las personas enfermas. Todos los móviles con la aplicación instalada se encargaban periódicamente de acceder al servidor para chequear si tenían en su memoria alguno de esos identificadores, lo que supondría el haber estado en contacto (en proximidad) con una persona enferma.

En el método centralizado, los identificadores que un móvil recoge de aquellos de los que estuvo próximo son enviados al servidor central. Cuando alguien da positivo, lo notifica al servidor central y es este el que se encarga directamente de notificar a todos aquellos móviles que hubiesen estado suficientemente próximos al de la persona enferma, y durante un tiempo suficiente, de esta circunstancia.

Fijémonos en que ni durante una pandemia se impuso el modelo centralizado, a pesar del valor incuestionable que podría tener para hacer estudios epidemiológicos, en particular los asociados a la forma y evolución de los contagios. Por eso, disponer de arquitecturas de datos y de aprendizaje automático que preserven su propiedad y, sobre todo, su privacidad, es cada vez más importante. El Reglamento General de Protección de Datos (RGPD) de la Unión Europea es paradigmático en este sentido. Cuando no existe una solución fácil o viable para reunir los datos, necesitamos otros enfoques distintos del aprendizaje centralizado. Eso explica por qué, aunque el aprendizaje federado se encuentra todavía en sus primeras etapas, en realidad en su infancia, está recibiendo mucha atención no solo por parte de la comunidad científica, sino también por parte de las empresas.

5. APRENDIZAJE FEDERADO: ORIGEN, FUNDAMENTOS Y EJEMPLOS

El aprendizaje federado (FL, *Federated Learning*) designa un conjunto de técnicas para entrenar modelos de aprendizaje automático cuando los datos no pueden o no deben centralizarse. La idea esencial es desplazar el cómputo hacia donde residen los datos (dispositivos u organizaciones) y agregar actualizaciones de los modelos aprendidos localmente (no los datos en bruto) para construir un modelo común, que pasa a ser compartido entre los aprendedores locales. Esta formulación reordena prioridades clásicas del aprendizaje automático: junto a la precisión del modelo que se aprende, pasan a primer plano la privacidad, la gobernanza del dato, el coste de comunicación y la robustez frente a la heterogeneidad, entre otras cuestiones menos relevantes en los sistemas centralizados.

El aprendizaje federado surge como respuesta a un problema cada vez más frecuente en sistemas modernos: los datos valiosos para entrenar modelos (texto tecleado en los móviles, sensores de dispositivos de vestir, historias clínicas o transacciones financieras, por ejemplo) son sensibles, voluminosos y, a menudo, están sujetos a restricciones legales y contractuales. Centralizarlos en una base de datos estructurada o incluso en un lago de datos (*data lake*) puede ser inviable o simplemente no recomendable. En ese contexto, la publicación seminal (McMahan *et al.,* 2017; McMahan *et al.,* 2016) que acuñó el término "Federated Learning" propone explícitamente entrenar modelos sin mover los datos, coordinando una "federación" de clientes —a los que a veces nos referimos como aprendedores— con un servidor que agrega actualizaciones de los modelos aprendidos localmente.

Los autores de este artículo que da inicio al campo del aprendizaje federado utilizan redes profundas (DNN, o *Deep Neural Networks*) como modelo de aprendizaje. Aunque este tipo de modelos es muy popular en el aprendizaje profundo, no es necesario ni mucho menos, ya que el aprendizaje federado es agnóstico al modelo o modelos de aprendizaje considerados —pueden aplicarse modelos de regresión logística, modelos lineales u otros enfoques—[2]. El utilizar redes profundas en este caso y en muchos otros puede facilitar el diseño de clasificadores y predictores con datos no IID (no independientes ni idénticamente distribuidos), especialmente complejos de tratar, como luego veremos. Por ejemplo, si varios hospitales colaboran para entrenar un modelo que prediga enfermedades, cada centro atiende a pacientes con perfiles distintos: algunos con mayor población anciana, otros con más jóvenes o con distintas patologías previas. Aunque todos midan lo mismo, los datos locales presentan distribuciones diferentes, lo que hace que el aprendizaje global sea más complicado.

La generación de un modelo global a partir de los modelos locales obtenidos por los aprendedores es relativamente sencilla y escalable. En este caso, los autores propusieron el ahora muy popular algoritmo *Federated Averaging* (FedAvg), una generalización del *stochastic gradient descent* para redes neuronales profundas entrenadas de forma distribuida.

En el apéndice 8 se incluye y explica el algoritmo original, FedAvg, por ser el propuesto para el caso del artículo fundacional del campo y ser, todavía, en sus diferentes versiones, ampliamente utilizado.

[2] En nuestro grupo de investigación hemos propuesto la arquitectura ECFL (*Ensemble and Continual Federated Learning*), pensada para escenarios más "reales" que los que puede abordar el aprendizaje federado clásico: múltiples dispositivos, datos en flujo continuo y cambios de distribución (*concept drift*), que luego analizaremos con detalle. La idea central es que, en lugar de intentar entrenar un único modelo global común (como en FedAvg), el modelo global sea un comité (*ensemble*) formado por varios modelos locales entrenados de manera independiente en cada cliente. Esto permite agregar modelos heterogéneos (distintas familias algorítmicas o estructuras), porque la agregación ya no requiere promediar parámetros, sino combinar predicciones (Casado *et al.,* 2023).

Pero el propio marco teórico del artículo deja claro que FL es agnóstico al modelo: en principio, puede aplicarse a regresión logística, modelos lineales u otros enfoques, siempre que el entrenamiento se base en actualizaciones de parámetros agregables.

Uno de los primeros casos de uso publicados sobre aprendizaje federado se orientó a la mejora de las sugerencias del teclado virtual Gboard de Google (Yang *et al.,* 2018). En concreto, el aprendizaje federado se usó para mejorar las sugerencias de consulta y de palabras (*query suggestions/next-word prediction*) que Gboard muestra mientras el usuario escribe. Cuando Gboard muestra una consulta sugerida a un usuario, su teléfono almacena localmente información sobre el contexto actual y si ha hecho clic en la sugerencia. El aprendizaje federado procesa ese historial en el dispositivo para sugerir mejoras a la siguiente iteración del modelo de sugerencia de consultas de Gboard. El objetivo era entrenar modelos de lenguaje a partir de lo que los usuarios teclean realmente, sin enviar ese texto —altamente sensible— a los servidores de Google.

La idea es aprender palabras fuera de vocabulario (OOV, *Out Of Vocabulary*) basándose en las palabras escritas con frecuencia por los usuarios de dispositivos móviles. OOV se refiere a palabras que no están incluidas en el vocabulario del dispositivo móvil de un usuario. Las palabras que faltan en el vocabulario no pueden predecirse mediante la sugerencia de teclado, la autocorrección o la escritura por gestos. Este es un claro ejemplo en el que no es posible compartir datos (palabras OOV) de los distintos usuarios (datos personales y sensibles), ni es factible aprender de la experiencia individual de cada usuario, porque es muy estrecha y extraordinariamente lenta. El aprendizaje federado es particularmente útil para resolver esta tarea, entrenando un modelo compartido de generación de OOV basado en los datos de todos los usuarios de móviles sin necesidad de transmitir datos sensibles a un servidor centralizado.

Durante este proceso, el modelo de generación de OOV en el móvil de cada usuario se actualiza constantemente, mientras que los datos de entrenamiento permanecen en el dispositivo. Como resultado, cada dispositivo móvil acaba teniendo un potente modelo de generación de OOV.

Otro ejemplo desarrollado, en este caso en nuestro grupo de investigación, consiste en haber usado el aprendizaje federado para el reconocimiento de la actividad de caminar a partir de los sensores inerciales de los móviles (Casado *et al.,* 2020), pero en condiciones realistas donde el móvil puede ir en mano, bolsillo, mochila, etc., y su orientación cambia constantemente. De nuevo se trata de una aplicación donde los datos recabados por los sensores de los móviles de los distintos usuarios son sensibles y privados.

En general, acudir a arquitecturas de aprendizaje federado responde al interés o necesidad de preservar la privacidad de datos obtenidos o distribuidos de múltiples fuentes, pero conviene subrayar que el aprendizaje federado no equivale automáticamente a una garantía plena de privacidad. El diseño original se apoya en la minimización de

datos (no subir datos brutos), pero el hecho de compartir actualizaciones de parámetros puede aportar información que permita en algunos casos recuperar datos o información de origen que siga siendo sensible (Suliman y Leith, 2023). Por eso muchas veces se incorporan medidas adicionales al aprendizaje federado, como el cifrado de datos o la agregación segura (Bonawitz *et al.,* 2017).

Creemos importante destacar que la contribución seminal al campo del aprendizaje federado se hizo por investigadores de Google. Pero aún más relevante es que Google no se limitó a proponer un algoritmo, sino que desarrolló y describió una arquitectura de sistema para ejecutar aprendizaje federado en producción con millones de dispositivos, abordando problemas que apenas aparecen en entornos de laboratorio (selección de clientes, fallos, sincronización por rondas, despliegue continuo, restricciones energéticas y de conectividad) (Bonawitz *et al.*, 2019). Este salto —del algoritmo a la plataforma operativa— es una de las razones por las que el aprendizaje federado se convirtió en un área propia y no en una simple variante del aprendizaje distribuido.

Hay que reconocer a las grandes compañías tecnológicas haber hecho algunas de las contribuciones más relevantes de los últimos años a las ciencias de la computación y a la IA en particular. De hecho, también es una aportación de Google la arquitectura de transformadores (Vaswani *et al.,* 2017), central en los LLM modernos. El paralelismo no es superficial: tanto en el aprendizaje federado como en los transformadores se observa un patrón recurrente de contribución "de plataforma".

La literatura sobre FL ha crecido de forma muy rápida desde 2017. Un estudio bibliométrico basado en Web of Science, en el que se analizan 3.107 artículos (Algorabi *et al.,* 2024) pone de manifiesto el rápido crecimiento en publicaciones en el campo, y el protagonismo de China, seguido a gran distancia por EE. UU. De hecho, entre las organizaciones con un mayor número de publicaciones en el período analizado (2017-2023), las diez primeras son chinas (Hong Kong incluido), si bien las principales revistas que recogen publicaciones de aprendizaje federado son del IEEE (por ejemplo, *IEEE Internet of Things Journal* e *IEEE Access*).

Tras el artículo seminal, el progreso en el campo se ha centrado más en ir realizando aportaciones algorítmicas y metodológicas, cuyo valor se mide más sobre problemas de laboratorio o conjuntos de datos abiertos pero sencillos (como MNIST), lo que está muy alejado de la solución a problemas reales de cierta complejidad. Esto denota que el campo todavía está más asentando conceptos, aportando recursos y herramientas computacionales y explorando las limitaciones de la tecnología de aprendizaje federado actual, que usándola en la resolución de problemas del mundo real. Sin duda, faltan algunos años para darle la suficiente madurez al campo.

De hecho, si analizamos algunos sectores como salud, finanzas o robótica personal e industrial, en los que el aprendizaje federado está llamado a tener un gran prota-

gonismo, la brecha entre expectativas y realidades es todavía muy grande. En salud, por ejemplo, gran parte de la evidencia publicada corresponde a escenarios *cross-silo* (hospitales o centros) y a menudo se limita a prototipos o estudios con un número reducido de instituciones, con desafíos persistentes de interoperabilidad y heterogeneidad clínica (Shah *et al.,* 2025). En finanzas, además de la heterogeneidad y la latencia, entran con fuerza cuestiones de auditoría, explicabilidad y cumplimiento normativo; la bibliografía reciente insiste en que la adopción práctica se ve condicionada por gobernanza, riesgos operacionales y amenazas de seguridad específicas (Kennedy *et al.*, 2025). En robótica y sistemas ciberfísicos, la dificultad crece porque el aprendizaje suele ser secuencial, sensible al tiempo real y frecuentemente requiere manejar gran cantidad de datos (en general procedentes de sensores que aportan información de la interacción de los robots con el mundo físico) y políticas de control con garantías: el aprendizaje federado encaja peor cuando la comunicación es cara y la distribución no estacionaria es la norma.

En conjunto, la conclusión razonable tras casi una década de desarrollo del aprendizaje federado es doble: (i) ha demostrado viabilidad operativa en algunos productos a gran escala (especialmente móviles); y (ii) su generalización a dominios "duros" y altamente regulados progresa, pero con un ritmo más lento y una dependencia fuerte de ingeniería, gobernanza y robustez.

6. QUEDA MÁS CAMINO QUE EL ANDADO

Como en cualquier nuevo campo científico-tecnológico, en el aprendizaje federado las preguntas son muchas más que las repuestas. Son muchos los problemas sin resolver que suponen barreras importantes a su aplicación generalizada a problemas del mundo real. Aunque no es una lista exhaustiva de los mismos, los problemas recurrentes y que explican por qué muchos éxitos siguen siendo "de laboratorio" o de alcance limitado, pueden agruparse en cinco frentes.

6.1. Comunicación, latencia y disponibilidad

Una de las limitaciones estructurales del aprendizaje federado es que el proceso de entrenamiento depende críticamente de la comunicación entre los participantes y el servidor coordinador. A diferencia del aprendizaje centralizado, donde el acceso a los datos es inmediato y continuo, en entornos federados la comunicación es costosa, limitada y heterogénea, tanto en ancho de banda como en fiabilidad (McMahan *et al.,* 2017; Bonawitz *et al.,* 2019).

En muchos escenarios reales —especialmente en configuraciones *cross-device*[3]— los participantes presentan disponibilidad intermitente. Dispositivos móviles pueden estar apagados, sin batería o sin conexión; robots o sistemas ciberfísicos pueden encontrarse fuera de servicio; y nodos industriales u hospitalarios pueden tener ventanas de comunicación restringidas por razones operativas o de seguridad (Yang *et al.,* 2018). Como consecuencia, en cada ronda federada solo una fracción de los participantes puede contribuir efectivamente al entrenamiento, lo que introduce variabilidad y ralentiza la convergencia.

Este problema se agrava cuando el modelo a entrenar es grande o complejo, como ocurre con redes neuronales profundas. En estos casos, cada ronda implica el intercambio de millones de parámetros, y la frecuencia de actualización necesaria para mantener un aprendizaje estable puede hacer que el coste de comunicación domine completamente el proceso. Incluso algoritmos diseñados para ser eficientes, como FedAvg, pueden volverse impracticables si el volumen de datos transmitidos y la latencia de sincronización superan los límites del sistema.

La latencia añade un segundo nivel de dificultad. El aprendizaje federado suele organizarse en rondas sincronizadas, lo que implica esperar a que un conjunto suficiente de participantes complete su entrenamiento local antes de agregar los resultados. En entornos con nodos lentos o poco fiables, este mecanismo introduce retrasos significativos o fuerza a descartar participantes, reduciendo la eficiencia estadística del aprendizaje. Existen alternativas asíncronas, pero introducen a su vez problemas de estabilidad y de consistencia del modelo global.

En conjunto, la combinación de coste de comunicación, latencia y disponibilidad variable no es un simple inconveniente de implementación, sino un factor que condiciona qué problemas pueden abordarse de forma realista con aprendizaje federado. Mientras estas limitaciones no se mitiguen de forma sistemática —mediante compresión, reducción de frecuencia de comunicación, arquitecturas jerárquicas o infraestructuras dedicadas—, el uso del aprendizaje federado en problemas complejos y a gran escala seguirá siendo selectivo y dependiente del contexto operativo.

6.2. Privacidad, seguridad y amenaza adversarial

Aunque el aprendizaje federado se presenta a menudo como una solución "intrínsecamente privada", en realidad introduce nuevos riesgos de seguridad y privacidad que no aparecen —o lo hacen de forma distinta— en el aprendizaje centralizado. El hecho de que los datos no se compartan directamente no implica que el sistema sea inmune a

[3] Suele usarse la referencia *cross-device* para hablar de sistemas con muchos dispositivos individuales, no coordinados, con alta variabilidad. Por el contrario, cuando son pocos participantes (organizaciones), estables y bien conectados se habla de *cross-silo*.

ataques: el intercambio de actualizaciones de modelo abre vectores de ataque específicos que deben considerarse explícitamente (Bonawitz *et al.,* 2017).

Uno de los problemas más estudiados es la corrupción o envenenamiento del modelo (*model poisoning*). En este tipo de ataque, uno o varios participantes maliciosos envían actualizaciones manipuladas con el objetivo de degradar el rendimiento global o forzar comportamientos indeseados (Suliman y Leith, 2023). Una variante especialmente peligrosa es la introducción de puertas traseras (*backdoors*), donde el modelo funciona correctamente en la mayoría de los casos, pero responde de forma errónea ante patrones específicos elegidos por el atacante. En entornos federados, detectar estos ataques resulta difícil, ya que no se tiene acceso a los datos locales y las actualizaciones pueden parecer estadísticamente plausibles.

Otro frente crítico es la integridad de la agregación. El servidor coordinador asume que las actualizaciones recibidas reflejan entrenamiento legítimo, pero en ausencia de mecanismos adicionales no puede verificar si los participantes son confiables ni la corrección de los gradientes. Incluso en escenarios sin participantes maliciosos, errores de implementación o fallos en dispositivos pueden introducir ruido sistemático que sesgue el modelo global.

Desde el punto de vista de la privacidad, se ha demostrado que las actualizaciones de modelo pueden filtrar información sensible. Los ataques de inferencia permiten, en ciertos casos, reconstruir características de los datos locales o inferir la pertenencia de un individuo a un conjunto de entrenamiento, a partir de gradientes o parámetros compartidos (Zhu *et al.,* 2019). Para mitigar estos riesgos se han propuesto técnicas como la agregación segura, que impide al servidor observar actualizaciones individuales, y la privacidad diferencial, que limita formalmente la información que puede inferirse sobre un participante concreto (Troncoso *et al*., 2022).

Sin embargo, estas defensas no son gratuitas. La agregación segura introduce sobrecoste computacional y de comunicación, además de complejidad criptográfica. La privacidad diferencial requiere añadir ruido a las actualizaciones o a los resultados agregados, lo que suele traducirse en una pérdida de precisión o en una necesidad mayor de datos y rondas de entrenamiento para alcanzar un rendimiento comparable. En problemas complejos, este compromiso entre privacidad, seguridad y utilidad se vuelve especialmente delicado.

En conjunto, la seguridad y la privacidad en aprendizaje federado no son propiedades automáticas, sino objetivos de diseño que deben equilibrarse cuidadosamente con la eficiencia y la calidad del modelo. La presencia de amenazas adversariales realistas implica que el aprendizaje federado, lejos de eliminar el problema de la confianza, lo

reformula: ya no se trata solo de confiar en una infraestructura central, sino de gestionar la desconfianza entre múltiples participantes y el propio coordinador del sistema.

6.3. Evaluación y validación clínica/regulatoria

En dominios de alto impacto social y económico, como la salud o las finanzas, el éxito de un sistema de aprendizaje automático no se mide únicamente por su rendimiento predictivo. En estos contextos es imprescindible cumplir requisitos adicionales de trazabilidad, validación externa, control de sesgos, reproducibilidad y auditoría, que están estrechamente ligados a marcos regulatorios y a la gestión del riesgo (Shah *et al.,* 2025). El aprendizaje federado introduce dificultades específicas para satisfacer estos requisitos, ya que rompe muchos de los supuestos habituales de evaluación en entornos centralizados.

Uno de los principales retos es la validación externa. En escenarios federados, los datos permanecen distribuidos entre múltiples instituciones o dispositivos, lo que dificulta la construcción de conjuntos de validación independientes y representativos. Cada nodo puede evaluar el modelo con sus propios datos, pero estas evaluaciones no son necesariamente comparables, ya que las distribuciones locales difieren y las métricas pueden reflejar realidades muy distintas. En salud, por ejemplo, un modelo puede funcionar bien en un hospital concreto y fallar sistemáticamente en otro con una población diferente, sin que exista un mecanismo sencillo para detectar este problema de forma global.

La trazabilidad y la reproducibilidad constituyen otro punto crítico. En un sistema federado, el modelo global es el resultado de múltiples rondas de entrenamiento con subconjuntos variables de participantes, datos que evolucionan en el tiempo y posibles actualizaciones asíncronas. Reconstruir exactamente cómo se obtuvo una versión concreta del modelo —qué nodos participaron, con qué datos y bajo qué condiciones— es mucho más complejo que en un entorno centralizado, lo que complica tanto la auditoría técnica como el cumplimiento normativo.

El control de sesgos se ve igualmente afectado. En ausencia de una visión global de los datos, identificar sesgos sistemáticos hacia determinados subgrupos resulta difícil. En sectores regulados, esta limitación es especialmente problemática, ya que las autoridades suelen exigir evidencias claras de equidad y ausencia de discriminación. El aprendizaje federado requiere, por tanto, nuevas estrategias para evaluar y mitigar sesgos sin acceder directamente a los datos subyacentes (He *et al.,* 2020).

Por último, la evaluación continua en producción plantea retos adicionales. En entornos clínicos o financieros, los modelos deben monitorizarse para detectar degradaciones de rendimiento o cambios de comportamiento que puedan tener consecuencias graves.

En un sistema federado, esta monitorización debe realizarse de forma distribuida y respetando las restricciones de privacidad, lo que ha motivado líneas de investigación específicas sobre evaluación federada, métricas agregadas seguras y protocolos de validación distribuidos (Kairouz *et al.,* 2021).

En conjunto, la dificultad de evaluar y validar modelos en aprendizaje federado no es un problema accesorio, sino un factor limitante clave para su adopción en dominios regulados. Superar esta barrera requiere no solo avances algorítmicos, sino también marcos metodológicos y organizativos que permitan garantizar confianza, cumplimiento y responsabilidad en sistemas donde los datos no pueden centralizarse.

6.4. MLOps federado y gobernanza del ciclo de vida

MLOps (*Machine Learning Operations*) es el conjunto de prácticas, procesos y herramientas que permiten desarrollar, desplegar, monitorizar y mantener modelos de aprendizaje automático en producción de forma fiable y reproducible. Por tanto, entrenar un modelo mediante aprendizaje automático —y federado, en particular— es solo una parte del problema.

Una síntesis útil es que el aprendizaje federado funciona mejor cuando coinciden tres condiciones que afectan a los datos, al problema y al sistema: muchos datos útiles pero no exportables; el problema a resolver debe tener un objetivo común claro entre los participantes —aprendedores—, de modo que todos optimizan esencialmente la misma tarea (por ejemplo, predicción de texto, detección de fraude, diagnóstico); una infraestructura capaz de orquestar el entrenamiento. El aprendizaje federado no es solo un algoritmo, sino un sistema distribuido complejo. Requiere seleccionar participantes, gestionar comunicación, tolerar fallos, asegurar privacidad, versionar modelos y monitorizar su evolución. Sin una infraestructura que coordine estas operaciones de forma robusta y automatizada, los costes operativos superan rápidamente los beneficios teóricos del enfoque.

Cuando falla alguna de estas condiciones, el beneficio puede diluirse frente a alternativas (aprendizaje centralizado con fuertes controles, enclaves seguros, intercambio de datos sintéticos o colaboración mediante modelos ya preentrenados).

6.5. Heterogeneidad no IID y desalineación de objetivos

Quizás este sea el problema más importante y complejo de resolver. En nuestro grupo de investigación está siendo uno de nuestros retos de investigación más relevantes. El problema surge cuando los datos entre clientes/instituciones difieren en distribución,

calidad, codificación y sesgos. Además, los objetivos pueden no estar alineados (por ejemplo, distintos hospitales con protocolos distintos). Esto tensiona tanto la convergencia como la equidad del rendimiento entre participantes, y obliga a técnicas de personalización y/o modelos por dominio que complican el despliegue.

Vamos a tratar de exponer la complejidad del problema.

6.5.1. Datos no independientes y no idénticamente distribuidos

Como ya hemos dicho, una de las diferencias fundamentales entre el aprendizaje automático clásico y el aprendizaje federado es que, en este último, los datos no se recogen ni se mezclan en un único repositorio, sino que proceden de múltiples fuentes autónomas (dispositivos, usuarios, instituciones). Esta característica rompe, en muchos casos, la hipótesis estadística habitual de que los datos son independientes e idénticamente distribuidos (IID), sobre la que se apoyan gran parte de los algoritmos de aprendizaje.

Para comprender las implicaciones de esta ruptura, conviene analizar por separado ambos conceptos: independencia e identidad de distribución, ya que describen propiedades distintas de los datos.

6.5.2. Independencia o no de los datos

Dos conjuntos de datos son independientes cuando la observación de uno no proporciona información sobre el otro. En el contexto del aprendizaje federado, esto se traduce en que los datos generados por una fuente (cliente o aprendedor) no influyen causalmente en los datos generados por otra.

Un ejemplo sencillo de datos independientes sería el de varios sensores de temperatura instalados en ubicaciones geográficas alejadas y sin interacción entre sí. Cada sensor genera sus mediciones de forma autónoma, y conocer las lecturas de uno no permite inferir las de otro, ni de forma aproximada. De manera similar, en un sistema federado con usuarios independientes que escriben textos privados en sus móviles, los datos de un usuario no afectan directamente a los de los demás.

Por el contrario, los datos no independientes aparecen cuando existe correlación o dependencia entre las fuentes. Un ejemplo típico es una red social, donde el comportamiento de un usuario influye en el de otros (mensajes, tendencias, reacciones). En un escenario federado, esto implica que los datos de distintos clientes están acoplados, lo que viola la hipótesis de independencia y complica tanto el entrenamiento como la evaluación de los modelos.

6.5.3. Distribución idéntica o no de los datos

Los datos son idénticamente distribuidos cuando todos proceden de la misma distribución estadística, incluso aunque sean independientes. En aprendizaje federado, esto significa que todos los clientes observan datos con características similares y en proporciones comparables.

Un ejemplo de datos idénticamente distribuidos sería un conjunto de dispositivos idénticos operando en condiciones controladas, como sensores industriales calibrados que miden la misma variable en procesos repetitivos. En este caso, cada cliente ve esencialmente el mismo tipo de datos, y los algoritmos de aprendizaje funcionan de forma similar a un entorno centralizado.

En cambio, los datos no idénticamente distribuidos son aquellos en los que cada fuente observa una distribución distinta. Este es el caso más común en aprendizaje federado real. Por ejemplo, distintos hospitales atienden a poblaciones de pacientes con características demográficas y clínicas diferentes; o distintos usuarios de móviles escriben en idiomas, registros y estilos distintos. Aunque los datos puedan ser independientes entre sí, sus distribuciones difieren, lo que introduce heterogeneidad estadística entre clientes.

6.5.4. Combinaciones habituales en aprendizaje federado

Al combinar ambas propiedades, se obtienen distintos escenarios relevantes en sistemas federados. El caso independiente e idénticamente distribuido (IID) se da cuando cada cliente o aprendedor recibe una muestra aleatoria de un mismo conjunto global de datos. Este escenario es común en experimentos controlados, pero poco representativo de aplicaciones reales.

Un escenario más realista es el de datos independientes pero no idénticamente distribuidos, donde los clientes no influyen entre sí, pero cada uno observa una distribución distinta. Este es el caso típico en aplicaciones como salud, finanzas o dispositivos personales, y constituye una de las principales dificultades del aprendizaje federado.

También pueden darse situaciones en las que los datos son no independientes, pero sí idénticamente distribuidos, como redes de sensores cercanos que miden el mismo fenómeno físico y presentan correlaciones espaciales o temporales.

Finalmente, el escenario más complejo corresponde a datos no independientes y no idénticamente distribuidos, donde existen tanto dependencias entre clientes como diferencias profundas en sus distribuciones: por ejemplo, en sistemas sociales, económicos o de tráfico.

En resumen, en aprendizaje federado, el término no IID engloba una variedad de situaciones muy distintas. Identificar correctamente si la falta de IID se debe a dependencias entre fuentes, a diferencias de distribución, o a ambas, es esencial para diseñar algoritmos adecuados y para interpretar de forma correcta los resultados obtenidos en escenarios reales.

6.5.5. Deriva de concepto

En los sistemas de aprendizaje federado que operan de forma continuada en el tiempo, una de las principales fuentes de degradación del rendimiento es la *deriva de concepto (concept drift)*. Este fenómeno hace referencia a cambios en el proceso generador de los datos que invalidan, total o parcialmente, las hipótesis bajo las cuales se entrenó el modelo. En contextos federados, la deriva de concepto adquiere una relevancia especial, ya que la adaptación del modelo es más costosa y compleja que en escenarios centralizados.

Desde un punto de vista probabilístico, cualquier problema de aprendizaje supervisado puede describirse mediante la distribución conjunta $P(x, y)$, que puede descomponerse como:

$$P(x, y) = P(x)P(y \mid x) \quad [5]$$

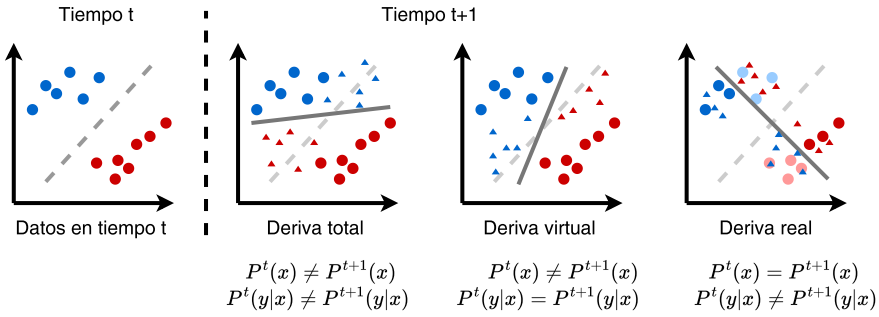
donde $P(x)$ representa la distribución de las entradas observadas y $P(y \mid x)$ la relación entre dichas entradas y las etiquetas. La deriva de concepto se produce cuando esta distribución conjunta cambia con el tiempo, es decir, cuando $P^t(x, y) \neq P^{t+1}(x, y)$. La utilidad de esta descomposición radica en que permite distinguir qué parte del problema está cambiando, lo cual tiene implicaciones directas sobre los mecanismos de detección y adaptación necesarios. En el estado inicial del sistema, correspondiente al instante t, los datos de entrada y las etiquetas mantienen una relación estable. Un modelo entrenado en estas condiciones puede aprender una frontera de decisión adecuada que separa correctamente las clases. Sin embargo, a medida que el sistema evoluciona y se recopilan nuevos datos en instantes posteriores, pueden aparecer distintos tipos de deriva.

Vamos a describir los distintos tipos de deriva, que gráficamente se muestran en la figura 5.

El caso más general y severo es la denominada *deriva total*, en la que cambian simultáneamente la distribución de las entradas y la relación entrada–salida:

FIGURA 5

CLASIFICACIÓN DE LA DERIVA DE CONCEPTO: COMPARATIVA ENTRE LA DERIVA TOTAL (CAMBIO EN ESPACIO DE ENTRADA Y EN LA RELACIÓN MODELO-ETIQUETA), VIRTUAL (CAMBIO SOLO EN EL ESPACIO DE ENTRADA) Y REAL (CAMBIO EN LA RELACIÓN MODELO-ETIQUETA)



Fuente: Elaboración propia.

$$P^t(x) \neq P^{t+1}(x), \qquad P^t(y \mid x) \neq P^{t+1}(y \mid x) \qquad [6]$$

En este escenario, los datos no solo se concentran en regiones distintas del espacio de entrada, sino que además el significado de las etiquetas respecto a esas entradas se ha modificado. Desde la perspectiva del aprendizaje federado, este es el caso más difícil de gestionar, ya que exige mecanismos de adaptación profunda del modelo y, habitualmente, la incorporación de nueva supervisión distribuida para evitar el deterioro progresivo del rendimiento.

Un caso más benigno es la *deriva virtual*, en la que únicamente cambia la distribución de las entradas, mientras que la relación entrada–salida permanece inalterada:

$$P^t(x) \neq P^{t+1}(x), \qquad P^t(y \mid x) = P^{t+1}(y \mid x) \qquad [7]$$

Aquí, el "concepto" subyacente no ha cambiado, pero los datos se presentan de forma distinta. En la práctica, este tipo de deriva puede estar causado por cambios en sensores —debido a su degradación, por ejemplo—, condiciones ambientales o patrones de uso. En sistemas federados, la deriva virtual suele poder abordarse mediante técnicas de adaptación más ligeras, como reentrenamientos parciales o ajustes en la ponderación de los clientes, sin necesidad de redefinir completamente el modelo.

Por último, la deriva real se produce cuando la distribución de las entradas se mantiene estable, pero cambia la relación entre entradas y etiquetas:

$$P^t(x) = P^{t+1}(x), \qquad P^t(y \mid x) = P^{t+1}(y \mid x) \qquad [8]$$

Este tipo de deriva es particularmente problemática, ya que no resulta evidente al analizar únicamente las entradas. El cambio afecta al significado de las clases o a la regla de decisión, lo que implica que un modelo entrenado previamente puede seguir observando datos "familiares", pero producir predicciones sistemáticamente erróneas. En aprendizaje federado, detectar y corregir una deriva real suele requerir acceso a nuevas etiquetas o señales de validación distribuidas, lo que introduce importantes retos operativos y de coordinación.

En conjunto, esta tipología pone de manifiesto que no toda deriva de concepto es equivalente. En sistemas federados y continuos, identificar si el cambio afecta a $P(x)$, a $P(y \mid x)$, o a ambos, es un paso esencial, y nada trivial, para diseñar estrategias de adaptación eficaces. Ignorar esta distinción conduce a soluciones parciales que funcionan en escenarios controlados, pero que fallan cuando el sistema se enfrenta a la complejidad y no estacionalidad propias de aplicaciones reales.

En nuestro grupo hemos abordado este problema de la no estacionalidad temporal de los datos (Casado *et al.*, 2022). Aunque FedAvg funciona razonablemente bien con heterogeneidad entre clientes, no está diseñado para adaptarse cuando la distribución de datos cambia con el tiempo, algo frecuente en dispositivos inteligentes (móviles, dispositivos de vestir, robots) y que provoca degradación de rendimiento si no se detecta y gestiona adecuadamente. En nuestro caso hemos propuesto el algoritmo CDA-FedAvg (*Concept-Drift-Aware Federated Averaging*), una extensión de FedAvg orientada a aprendizaje federado continuo en un único objetivo compartido (*single-task*), pero con posibles cambios de distribución en el tiempo. A diferencia de FedAvg "por rondas" estrictas, planteamos un esquema asíncrono en el que los clientes ganan autonomía para decidir cuándo entrenar y qué datos usar, mientras el servidor actúa sobre todo como orquestador y agregador cuando recibe actualizaciones.

Por otra parte, en 2022 publicamos un artículo de revisión (Criado *et al*., 2022) cuyo objetivo era ordenar y clarificar la heterogeneidad estadística de los datos no IID, tanto entre clientes (cada dispositivo/institución tiene datos distintos) como a lo largo del tiempo (los datos cambian, y nos enfrentamos al *concept drift*). Evidenciamos que la mayor parte de las soluciones obviaban el problema del no IID y que, a menudo, cuando lo hacían, no precisaban con suficiente cuidado qué tipo de heterogeneidad se asumía, lo que dificulta comparar métodos y entender cuándo funcionan realmente.

La contribución central de nuestro trabajo fue conceptual y taxonómica: proponiendo una clasificación formal de la heterogeneidad y revisando estrategias de aprendizaje federado destacadas en función de esa clasificación (por ejemplo, métodos orientados a mitigar la

divergencia por datos no IID, técnicas de personalización, ajustes en agregación/optimización, etc.). En paralelo, el artículo introduce de forma explícita el puente con el aprendizaje continuo (*Continual Learning* o CL): subraya que muchas situaciones federadas reales no solo son "no IID entre clientes", sino también no estacionarias en el tiempo, y que herramientas clásicas de aprendizaje continuo (por ejemplo, mecanismos para manejar deriva y reducir olvido) pueden adaptarse al entorno federado para mejorar robustez.

La tabla 1 organiza los distintos tipos de heterogeneidad de datos en aprendizaje federado, atendiendo a dos ejes conceptuales ortogonales:

- Heterogeneidad espacial, es decir, diferencias estadísticas entre los datos de distintos participantes en un mismo instante, y
- Heterogeneidad temporal, es decir, cambios en la distribución de los datos a lo largo del tiempo.

TABLA 1

TAXONOMÍA DE LA HETEROGENEIDAD ESPACIAL Y TEMPORAL EN APRENDIZAJE FEDERADO.

		Heterogeneidad espacial			
		Sin cambios (datos IID)	Cambios en el espacio de entrada entre clientes	Cambios en el comportamiento entre clientes	Cambios entrada y comportamiento entre clientes
Heterogeneidad temporal	Sin cambios (datos IID)				
	Deriva de concepto virtual		NO ABORDADO HASTA AHORA		
	Deriva de concepto real				
	Deriva total de concepto				

Nota: La zona sombreada indica combinaciones de heterogeneidad que no han sido abordadas de forma efectiva por los métodos existentes. En el eje horizontal se muestra la heterogeneidad espacial, es decir, diferencias estadísticas entre los datos de distintos participantes en un mismo instante, y en el eje vertical, heterogeneidad temporal, es decir, cambios en la distribución de los datos a lo largo del tiempo.

En el eje horizontal se representan los distintos grados de heterogeneidad espacial, desde el caso ideal de datos IID hasta escenarios en los que existen diferencias entre

clientes en el espacio de entrada $P_i(x)$, en el comportamiento o relación entrada–salida $P_i(y \mid x)$ o en ambos simultáneamente. En el eje vertical se representan los distintos grados de heterogeneidad temporal, desde datos estacionarios hasta situaciones de deriva de concepto, distinguiendo entre deriva virtual (cambios en), deriva real (cambios en) y deriva total. En el eje horizontal se representan los distintos grados de heterogeneidad espacial, desde el caso ideal de datos IID hasta escenarios en los que existen diferencias entre clientes en el espacio de entrada $P_i(x)$, en el comportamiento o relación entrada–salida $P_i(y \mid x)$ o en ambos simultáneamente. En el eje vertical se representan los distintos grados de heterogeneidad temporal, desde datos estacionarios hasta situaciones de deriva de concepto, distinguiendo entre deriva virtual (cambios en $P(x)$), deriva real (cambios en $P(y \mid x)$) y deriva total.

Las celdas sombreadas indican regiones del espacio del problema que, hasta ahora, no han sido abordadas de forma satisfactoria por la literatura en aprendizaje federado. En particular, cuando coexisten heterogeneidad espacial en el comportamiento y deriva temporal, los métodos actuales muestran limitaciones claras, ya sea por problemas de convergencia, degradación del rendimiento o falta de mecanismos de adaptación. La tabla pone de manifiesto que buena parte de la investigación se ha centrado en escenarios parciales (por ejemplo, no IID espacial sin deriva temporal), mientras que los escenarios más realistas —donde ambas heterogeneidades coexisten— siguen siendo en gran medida un reto abierto.

La tabla 2 amplía el análisis anterior incorporando un elemento clave para la aplicabilidad real del aprendizaje federado: las restricciones prácticas sobre la disponibilidad de datos etiquetados. Manteniendo los mismos ejes de heterogeneidad espacial y temporal que en la tabla anterior, esta tabla indica qué supuestos adicionales son necesarios para que los métodos actuales puedan funcionar en cada escenario.

Las leyendas distinguen entre cuatro situaciones: ausencia de restricciones, disponibilidad de datos etiquetados de un solo participante en múltiples momentos (1P-MT), disponibilidad de datos etiquetados de todos los participantes en un único momento inicial (TP-1T) y disponibilidad de datos etiquetados de todos los participantes de forma periódica (TP-MT). Estas restricciones no describen algoritmos concretos, sino hipótesis implícitas que muchos trabajos asumen —a menudo de forma no explícita— para poder manejar la heterogeneidad y la deriva.

La tabla muestra que, a medida que aumenta la complejidad del escenario (heterogeneidad espacial combinada con deriva temporal), los métodos existentes requieren supuestos de supervisión cada vez más fuertes, como la necesidad de etiquetado periódico o la existencia de un participante "ancla" que proporcione datos de referencia. Esto pone de relieve una brecha importante entre los entornos experimentales y los despliegues reales, donde dichas condiciones suelen ser costosas, difíciles de mantener o directamente inviables.

TABLA 2

RESTRICCIONES PRÁCTICAS SOBRE LA DISPONIBILIDAD DE DATOS ETIQUETADOS NECESARIOS PARA ABORDAR DISTINTOS ESCENARIOS DE HETEROGENEIDAD ESPACIAL Y TEMPORAL EN APRENDIZAJE FEDERADO Y CONTINUO

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="4">Heterogeneidad espacial</td></tr>
<tr><td>Sin cambios (datos IID)</td><td>Cambios en el espacio de entrada entre clientes</td><td>Cambios en el comportamiento entre clientes</td><td>Cambios entrada y comportamiento entre clientes</td></tr>
<tr><td rowspan="4">Heterogeneidad temporal</td><td>Sin cambios (datos IID)</td><td></td><td></td><td colspan="2" rowspan="2">Suficientes datos etiquetados de cada participante, al inicio. Restricción TP-1T</td></tr>
<tr><td>Deriva de concepto virtual</td><td></td><td></td></tr>
<tr><td>Deriva de concepto real</td><td colspan="2" rowspan="2">Suficientes datos etiquetados de un participante, de forma periódica. Restricción 1P-MT</td><td colspan="2" rowspan="2">Suficientes datos etiquetados de cada participante, de forma periódica. Restricción TP-MT</td></tr>
<tr><td>Deriva total de concepto</td></tr>
</table>

Nota: En el eje horizontal se muestra la heterogeneidad espacial, es decir, diferencias estadísticas entre los datos de distintos participantes en un mismo instante, y en el eje vertical, heterogeneidad temporal, es decir, cambios en la distribución de los datos a lo largo del tiempo.

En conjunto, la tabla sugiere que el cuello de botella no es solo algorítmico, sino también organizativo y operativo: muchos enfoques actuales funcionan únicamente si se acepta un coste elevado en términos de etiquetado y coordinación entre participantes.

6.5.6. Datos ruidosos y de calidad desigual en los clientes

Además de las diferencias en la distribución de los datos entre clientes, los sistemas de aprendizaje federado reales suelen presentar heterogeneidad en los tipos de datos o en la calidad de la información aportada por cada cliente o aprendedor. No todos los clientes obtienen sus datos en las mismas condiciones ni con el mismo nivel de fiabilidad.

Por ejemplo, en una red de sensores o dispositivos inteligentes, algunos nodos pueden estar correctamente calibrados y operar de forma estable, mientras que otros pueden sufrir fallos intermitentes, ruido elevado o condiciones de operación singulares. En aplicaciones centradas en usuarios, normalmente hay participantes que generan datos abundantes y consistentes, mientras que otros contribuyen de forma esporádica o con patrones muy particulares. Aunque estas situaciones no impliquen necesariamente un comportamiento malicioso, su efecto sobre el aprendizaje conjunto puede ser similar.

En el aprendizaje federado clásico, el objetivo es entrenar un único modelo global que funcione razonablemente bien para todos los clientes. Bajo este enfoque, se asume implícitamente que las contribuciones de los distintos participantes son compatibles entre sí y que las diferencias observadas reflejan la variabilidad razonable del entorno o del uso del sistema. Este supuesto, aunque útil desde un punto de vista conceptual, resulta limitado cuando algunos clientes generan datos muy ruidosos, poco representativos o sistemáticamente inconsistentes.

El aprendizaje federado personalizado surge como una extensión natural de este planteamiento. En lugar de buscar una única solución común, estos métodos permiten que cada cliente disponga de un modelo parcialmente adaptado a sus características locales. Esta distinción es clave: mientras que el aprendizaje federado estándar prioriza la generalización global, el aprendizaje federado personalizado pone el énfasis en capturar la diversidad entre clientes sin perjudicar el rendimiento del conjunto.

Un reto fundamental en este contexto es que, debido a las restricciones de privacidad inherentes al aprendizaje federado, no es posible inspeccionar directamente los datos de los clientes para evaluar su compatibilidad o calidad. El servidor central no puede saber si dos clientes observan fenómenos similares, si sus datos son coherentes entre sí o si uno de ellos introduce ruido de forma sistemática. Esta limitación obliga a diseñar mecanismos indirectos que permitan inferir estas propiedades a partir del comportamiento del proceso de entrenamiento, sin violar la privacidad de los datos locales.

Uno de los trabajos desarrollados en nuestro grupo de investigación (Burés *et al.*, 2025) aborda este problema analizando cómo contribuyen los distintos clientes al aprendizaje conjunto a lo largo del tiempo, con el objetivo de identificar patrones consistentes y discrepantes. De forma intuitiva, el sistema aprende a diferenciar entre clientes cuyas aportaciones resultan compatibles con el objetivo común y aquellos cuya información introduce inestabilidad. Esto permite ajustar su influencia sin necesidad de acceder a los datos originales ni excluir explícitamente a ningún participante.

Los estudios realizados en entornos de simulación muestran que este enfoque conduce a mejoras consistentes en el rendimiento y la estabilidad del aprendizaje, especialmente en escenarios con datos heterogéneos, presencia de ruido o clientes con objetivos maliciosos. Estos resultados ponen de manifiesto la importancia de tratar explícitamente la compatibilidad entre clientes y refuerzan la idea de que la personalización y la robustez son componentes clave para el despliegue efectivo del aprendizaje federado en aplicaciones reales.

7. MENSAJE FINAL

El mensaje final —coherente con el subtítulo de nuestro trabajo de 2022: "A long road ahead"— es deliberadamente crítico: aunque ha habido avances significativos en estos años en el aprendizaje federado, identificamos importantes retos abiertos persistentes en condiciones realistas (heterogeneidad múltiple, evolución temporal, restricciones de comunicación, etc.). El ámbito del aprendizaje federado necesita separar la paja cada vez más abundante del grano, bases de datos de evaluación y mecanismos de validación más rigurosos y representativos de la complejidad real de las aplicaciones realmente útiles y, por supuesto, conseguir avances claros en todos los aspectos antes apuntados, y que suponen de facto una seria limitación a la resolución de problemas del mundo real mediante aprendizaje automático.

En todo caso, es incontestable que el aprendizaje federado supone una gran oportunidad para empresas y organizaciones de todo tipo al aportar privacidad a los datos, poder mejorar colectivamente el aprendizaje de múltiples aprendedores, con o sin personalización de lo aprendido, y otras ventajas que hemos ido comentando a lo largo de este texto. Sectores como salud, banca, movilidad, industria 4.0 o los servicios digitales pueden beneficiarse de modelos que aprendan de múltiples fuentes de manera coordinada y segura, reduciendo costes, mejorando la calidad de los servicios y acelerando la innovación. En este sentido, aunque el camino sigue siendo largo, los avances recientes y las estrategias emergentes de operación en contextos no IID, la personalización y la mayor robustez y seguridad de las nuevas arquitecturas, muestran que el aprendizaje federado puede convertirse en un recurso indispensable en el mundo empresarial y tecnológico.

Referencias

Algorabi, Ö. *et al.* (2024). A Bibliometric Analysis on Federated Learning. *Journal of Advanced Research in Natural and Applied Sciences*, 10, 875–898.

Bonawitz, K. *et al.* (2017). Practical Secure Aggregation for Privacy-Preserving Machine Learning. En: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, (1175–1191).

Bonawitz, K. *et al.* (2019). Towards Federated Learning at Scale: System Design. *arXiv*:1902.01046.

Burés, J. *et al.* (2025). Out of Distribution Detection and Adaptive Interpolation for Personalized Federated Learning. *Frontiers in Artificial Intelligence and Applications* (1873–1879).

Casado, F. *et al.* (2023). Ensemble and continual federated learning for classification tasks. *Machine Learning*, 112, 1–41.

Casado, F. E. *et al.* (2020). Walking Recognition in Mobile Devices. *Sensors*, *20*(4), art. 1189.

Casado, F. E. *et al.* (2022). Concept drift detection and adaptation for federated and continual learning. *Multimedia Tools Appl.*, 81(3), 3397–3419.

Criado, M. F. *et al.* (2022). Non-IID data and Continual Learning processes in Federated Learning: A long road ahead. *Information Fusion*, 88, 263-280.

Fernández-Delgado, M. *et al.* (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15, 3133–3181.

Fernández-Delgado, M. *et al.* (2019). An extensive experimental survey of regression methods. *Neural Networks*, 111, 11-34.

He, C. *et al.* (2020). FedML: A Research Library and Benchmark for Federated Machine Learning. *arXiv:2007.02945*.

Hinton, G. E., Osindero, S., y Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.*, *18*(7), 1527–1554.

Kairouz, P. *et al.* (2021). Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, *14*(1–2), 1–210.

Kennedy, C., Hilal, A., y Momeni, M. (2025). The Role of Federated Learning in Improving Financial Security: A Survey. En: *GCAIoT 2025* (1-8).

Krizhevsky, A., Sutskever, I., y Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. En: *Proceedings of NIPS'12* (1097–1105).

LeCun, Y. *et al.* (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., Yere, Y., y Hinton, G. (2015). Deep Learning. *Nature*, 521, 436-444.

McCarthy, J. (1959). Programs with Common Sense. En: *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (75–91). London.

McMahan, B. *et al.* (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. En: *Proceedings of AISTATS* (1273–1282).

McMahan, H. B. *et al.* (2016). Federated Learning of Deep Networks using Model Averaging. *CoRR*, abs/1602.05629.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-hill.

Mnih, V. *et al.* (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Newell, A., y Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM, 19*(3), 113–126.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Rumelhart, D. E., Hinton, G. E., y Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development, 3*(3), 210–229.

Shah, S. T. *et al.* (2025). Federated Learning in Public Health: A Systematic Review. *Healthcare*, *13*(21), art. 2760.

Shortliffe, E. (1976). Computer-based medical consultations: MYCIN. *Artificial Intelligence - AI*, 388, 1–20.

Silver, D. *et al.* (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, *529*(7587), 484–489.

Suliman, M., y Leith, D. (2023). Two Models are Better Than One. En: *Proceedings of ESORICS 2023* (105–122). Berlin: Springer.

Sutton, R. S., y Barto, A. G. (2018). *Reinforcement learning: an introduction*. Cambridge, MA: The MIT Press.

Troncoso, C. *et al.* (2022). Deploying decentralized, privacy-preserving proximity tracing. *Communications of the ACM*, 65, 48–57.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.

Vaswani, A. *et al.* (2017). Attention is all you need. En: *Advances in Neural Information Processing Systems*, (5998–6008).

YANG, T. *ET AL.* (2018). Applied Federated Learning: Improving Google Keyboard Query Suggestions. *CoRR*, abs/1812.02903.

ZHU, L., LIU, Z., Y HAN, S. (2019). Deep Leakage from Gradients. En: *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

APÉNDICE A. ALGORITMO FEDAVG

Algoritmo 1: Federated Averaging

Inicializar el modelo global w_0

Para cada ronda $t = 1, 2, \cdots, T$

  El servidor selecciona un subconjunto de clientes S_t

  Para cada cliente $k \in S_t$ (en paralelo):

    $w_{t+1}^k \leftarrow$ ClienteActualiza (k, w_t)

  El servidor agrega:

  $w_{t+1} \leftarrow \sum_k \frac{n_k}{n} \cdot w_{t+1}^k$

Función ClienteActualiza (k, w)

  $w_{local} \leftarrow w$;

  Para cada época local $i = 1, \cdots, E$:

    Para cada minibatch $b \in D_k$:

      $w_{local} \leftarrow w_{local} - \eta \cdot \nabla \ell(w_{local}; b)$

  Devuelvor w_{local} ;

Notación:

- w_t : parámetros del modelo global en la ronda t
- S_t : subconjunto de clientes seleccionados en la ronda t
- D_k : datos locales del cliente k

- n_k : número de ejemplos en D_k
- $n = \sum_k n_K$: total de ejemplos de los clientes participantes
- E : número de épocas locales
- η : tasa de aprendizaje
- ℓ : función de pérdida

FedAvg es una generalización del SGD distribuido (*Distributed Stochastic Gradient Descent*, es una extensión del descenso de gradiente estocástico en la que el entrenamiento de un modelo se reparte entre varios nodos de cómputo que procesan datos en paralelo), adaptada al escenario federado, donde los datos están distribuidos y no se comparten. La idea es la siguiente:

1. *Inicialización global*: El servidor mantiene un modelo global común a todos los clientes.

2. *Selección de clientes*: En cada ronda solo participa una fracción de clientes, reflejando disponibilidad intermitente (por ejemplo, móviles conectados).

3. *Entrenamiento local*: Cada cliente entrena el modelo sobre sus propios datos, realizando varias épocas de SGD local. Este paso reduce drásticamente la comunicación, ya que no se envían gradientes en cada *minibatch*.

4. *Agregación ponderada*: El servidor promedia los modelos locales, ponderándolos por el número de datos de cada cliente. El resultado es el nuevo modelo global.

CAPÍTULO III

Alfabetización en IA: hacia un uso responsable y controlado de la tecnología

Francisco Bellas*

Este capítulo analiza el impacto de la inteligencia artificial (IA) desde la perspectiva de la alfabetización, como condición necesaria para un uso responsable, seguro y socialmente beneficioso de esta tecnología. Tras describir el contexto de adopción acelerada de la IA y los riesgos asociados a su autonomía, opacidad y carácter probabilístico, se argumenta la necesidad de una formación que integre capacitación técnica, pensamiento crítico, responsabilidad legal y control humano. La parte final del capítulo plantea una reflexión sobre el impacto de la IA en el aprendizaje y, a modo de ejemplo, desarrolla el posible uso de la IA generativa como soporte para el estudio.

Palabras clave: alfabetización en IA, capacitación digital, pensamiento crítico, control humano, IA y aprendizaje.

* El CITIC, como centro con acreditación de excelencia del Sistema Universitario de Galicia y miembro de la Red CIGUS, está subvencionado por la Conselleria de Educación, Ciencia, Universidades y Formación Profesional de la Xunta de Galicia y cofinanciado a través del Fondo Europeo de Desarrollo Regional (FEDER), Programa operativo FEDER Galicia 2021-27 (Ref. ED431G 2023/01).

1. EL IMPACTO DE LA INTELIGENCIA ARTIFICIAL

Para los que llevamos años trabajando en inteligencia artificial (IA), ha resultado especialmente llamativo el auge que ha tenido este campo desde el lanzamiento de ChatGPT 3.5 en el año 2022 (OpenAI, 2022). De ser un área de investigación y desarrollo con grandes expectativas, pero con un avance progresivo y bastante especializado, ha pasado a protagonizar una auténtica revolución tecnológica que afecta ya a la mayoría de los segmentos de la sociedad. Es importante aclarar que esta revolución se fundamenta, por ahora, en una rama concreta de la IA, la IA generativa (Corredera, 2023), pero su empuje está afectando ya al progreso de otras áreas del campo.

¿Qué ha motivado este progreso tan acelerado? Principalmente, que la tecnología funciona, y cada vez mejor. Esta IA generativa es de gran utilidad en una multitud de tareas que hacen uso de herramientas digitales que manipulan datos. Desde aplicaciones ofimáticas, pasando por creación digital de imagen, audio o vídeo, para llegar a casos más avanzados, como sistemas de recomendación, modelos predictivos o herramientas de automatización de tareas. Cualquier empresa que utilice este tipo de herramientas digitales, y son mayoría, sabe que debe abrazar la IA generativa si quiere mantener la competitividad (World Economic Forum, 2025), ya que aumenta la eficiencia en multitud de trabajos.

Por otro lado, entre ocio y uso personal, encontramos cada vez más IA en dispositivos de uso cotidiano como teléfonos móviles, ordenadores, plataformas de *streaming*, electrodomésticos o vehículos, ya que mejora la experiencia del usuario al permitir una comunicación más natural (habla), automatizar tareas tediosas, proponer soluciones a problemas complejos y, por qué no decirlo, crear nuevas formas de divertimento (ENAE International Business School, 2025).

En resumen, tenemos una tecnología útil para la población general en el ámbito laboral y personal, y que encaja de manera directa en el ecosistema digital existente. Es decir, estamos ante una oportunidad de negocio muy lucrativa (Appio y Schiavone, 2023). Las grandes empresas tecnológicas han encontrado en la IA un filón inesperado, y están realizando grandes inversiones para su progreso. Es fácil imaginar la dependencia que esta tecnología puede crear en la población, y por tanto, su impacto económico futuro (Singla y Hall, 2025).

La adopción masiva de la IA ha dejado a la vista un conjunto de riesgos que no aparecían en tecnologías digitales anteriores. Por una parte, la IA puede resolver problemas muy complejos, dándonos respuestas también complejas, que podríamos no entender. Por otra parte, se diseña para tomar decisiones de manera autónoma, con mínima intervención humana. Finalmente, la IA basada en datos que domina las aplicaciones actuales se basa siempre en probabilidades de acierto, no en certezas, por lo que la incertidumbre en la respuesta es parte de su funcionamiento normal, incluyendo resul-

tados sesgados o inexactos. Es decir, estamos usando herramientas en nuestro nombre, sin comprender bien lo que hacen y sin estar totalmente seguros de que lo hagan correctamente, simplemente porque aparentan ser "inteligentes" (Bellas, 2025).

En consecuencia, nos encontramos actualmente en un escenario donde el modelo económico empuja el avance de la IA, pero donde es cada vez más evidente que se requiere un mayor control. En este sentido, Europa lleva la iniciativa con su Ley de IA (European Union, 2024), que rige la creación y el despliegue de esta tecnología con un claro foco en el respeto de los derechos básicos de los ciudadanos. De hecho, podríamos afirmar que el sector social más afectado actualmente por la IA en Europa es el legislativo, ya que debe avanzar en la aplicación de esta ley, mientras no se pone freno al ritmo de desarrollo de las empresas tecnológicas que comercializan IA (Agencia Española de Supervisión de Inteligencia Artificial —AESIA—, 2025).

En este contexto de tensión entre economía y control, la alfabetización en IA emerge como una condición necesaria para garantizar un uso *responsable, seguro* y socialmente *beneficioso* de esta tecnología. La IA introduce desafíos específicos derivados de su capacidad de decisión autónoma, su opacidad técnica y su impacto directo sobre procesos cognitivos humanos. Pero la historia de las transformaciones tecnológicas muestra que la educación ha sido siempre el principal mecanismo de adaptación social frente al cambio, ya que una población formada es más libre, menos manipulable, y esta capacidad es básica en el contexto de la IA (Bellas, 2024).

Este capítulo aborda la necesidad de una alfabetización en IA rigurosa, desde una perspectiva sistémica, conectando educación, regulación, empleo y ciudadanía digital. No es el objetivo aquí proponer soluciones concretas, ya que estas requieren de investigación y evidencia práctica, sino llamar la atención de los lectores sobre qué implica alfabetizar en IA y por qué se necesita comenzar lo antes posible.

2. ALFABETIZACIÓN EN IA

Existen diversas iniciativas internacionales para desarrollar alfabetizaciones en IA, principalmente a nivel preuniversitario, y con enfoques similares. En Europa, la iniciativa más destacada es el marco AILit (*AI Literacy Framework*), resultado de una colaboración entre la Organización para la Cooperación y el Desarrollo Económicos (OCDE) y la Comisión Europea (European Commission y OECD, 2025), lanzado en 2025 con el apoyo de Code.org y expertos internacionales. Este marco define 22 competencias distribuidas en cuatro dominios principales (comprensión, creación, gestión y diseño de IA) y constituirá la base de la evaluación PISA 2029 sobre Alfabetización en Medios e IA (OECD, 2025). De manera complementaria, la Unesco lanzó en 2024 dos marcos de competencias, uno para estudiantes y otro para docentes, con énfasis en un enfoque centrado en el humano que integra ética, pensamiento crítico y responsabilidad social (Unesco, 2024). En el contexto

de la Ley de IA, existe un artículo específico dedicado a la alfabetización en IA que será comentado más adelante, el cual ha generado un repositorio comunitario con más de 40 iniciativas prácticas implementadas por empresas e instituciones públicas (European Union, 2025).

En América del Norte, Estados Unidos ha dado un paso decisivo con la Orden Ejecutiva Presidencial de abril de 2025 (The White House, 2025), la cual establece un enfoque integral que incluye la creación de un grupo de trabajo específico de la Casa Blanca para coordinar esfuerzos federales en educación en IA, el desarrollo de estrategias concretas a nivel nacional para destacar logros de estudiantes y educadores en IA, y la implementación de asociaciones público-privadas para crear recursos educativos de alfabetización en IA y pensamiento crítico. Por otro lado, Canadá está promoviendo un diálogo nacional sobre IA responsable mediante su iniciativa "Learning Together" (Innovation, 2025), aunque aún carece de una estrategia nacional coordinada. En América Latina, Brasil ha emergido como protagonista con la iniciativa "ConectAI", una colaboración público-privada que ya ha capacitado a 2,5 millones de estudiantes en el uso de IA, con el objetivo de llegar a cinco millones para 2027 (Microsoft y Brazilian Government, 2025). Pero el resto de la región presenta un panorama más fragmentado, con 26 iniciativas identificadas, principalmente impulsadas por el sector privado (ProFuturo y OEI, 2025).

En Asia-Pacífico, China ha adoptado una posición de liderazgo hace años ya, haciendo obligatoria la educación en IA desde 2025 en todas sus escuelas primarias y secundarias, estableciendo incluso un mínimo de horas anuales de formación (Ministry of Education of China, 2025). Singapur ha desarrollado la plataforma "Learn AI Singapore" integrada en su estrategia nacional de IA, incorporando herramientas de aprendizaje en el sistema educativo público (AI Singapore, 2025). Australia ha establecido su propio marco centrado en seis principios fundamentales: enseñanza y aprendizaje, bienestar, transparencia, equidad, responsabilidad y privacidad, que se alinea con su currículo nacional sobre IA (Australian Education Council, 2025).

Tras haber quedado patente el interés a nivel internacional en la creación de alfabetizaciones, pasaré a continuación a detallar cómo se está abordando esta alfabetización en el caso concreto de Europa, del cual derivarán los futuros planes de estudio en España.

2.1. ¿Qué objetivo tiene?

El artículo 4 de la Ley de IA[1] entró en vigor en febrero de 2025. Exige a los proveedores y responsables del despliegue de sistemas de IA que garanticen un nivel suficiente de alfabetización en materia de IA de su personal y de otras personas que tratan con sistemas de IA en su nombre. Esto afecta a todos los que ponen herramientas de IA a

[1] https://artificialintelligenceact.eu/es/article/4/

disposición de sus empleados o alumnos. Estamos, por tanto, ante un requerimiento legal que debemos abordar en todos los Estados miembros.

Las recientes guías desarrolladas en la propia Unión Europea (UE) definen esta alfabetización en IA como "los conocimientos técnicos, las habilidades duraderas y actitudes orientadas al futuro necesarias para desenvolverse con éxito en un mundo influenciado por la IA" (European Commission y OECD, 2025). En este enfoque queda claro que formarse en esta tecnología va más allá de la educación digital tradicional (técnicas y habilidades) y requiere también actitudes (impacto de la IA). Se unen, por tanto, los enfoques tecnológico y social en una disciplina que parecía principalmente técnica.

Una alfabetización rigurosa en IA debe perseguir tres objetivos fundamentales (ver figura 1):

FIGURA 1

LOS TRES OBJETIVOS PRINCIPALES DE LA ALFABETIZACIÓN EN IA

Fuente: Imagen creada con IA generativa.

- *Capacitación*: El primero es capacitar a las personas en el uso competente de herramientas y sistemas de IA, para explotar todas sus ventajas. Este es el más familiar para los educadores, porque es lo que se hace con cualquier tecnología: enseñar a los estudiantes a utilizarla efectivamente. Aquí tenemos experiencia y enfoques pedagógicos bien establecidos dentro del área de la educación digital.

- *Responsabilidad*: El segundo objetivo, más específico de la IA, es desarrollar en los usuarios una comprensión clara de su responsabilidad cuando utilizan

estas herramientas. Esto incluye entender cómo funciona la IA, cuáles son sus limitaciones, cuándo debería o no confiarse en sus respuestas, y qué consecuencias pueden tener las decisiones informadas por sistemas de IA. Estamos hablando aquí de la dimensión ética de la IA, y también de su dimensión legal (Unesco, 2021).

- *Control*: El tercero es asegurar que las personas mantengan el control sobre su propio trabajo y sus propias vidas cuando utilizan la IA (*human oversight* en inglés). La supervisión y la capacidad de decisión individual deben estar siempre en el centro del uso de la IA, siendo esta un asistente para apoyar el trabajo humano. Fomentamos, pues, el pensamiento crítico de los usuarios de esta tecnología (Comisión Europea, 2025).

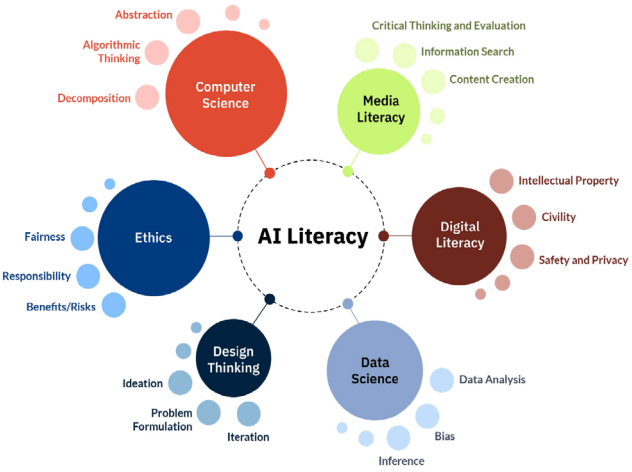
Logrando estos tres objetivos, tendremos una población capaz de hacer un uso eficiente y seguro de las herramientas de IA, manteniendo una opinión propia.

2.2. ¿Qué contenidos debe incluir?

La IA es un campo multidisciplinar. No se puede enfocar su aprendizaje únicamente como una rama de la informática. De hecho, el último marco de trabajo de la Unión Europea establece que una alfabetización en IA debe contener seis áreas básicas (European Commission y OECD, 2025), que se muestran en la figura 2:

FIGURA 2

MARCO DE ALFABETIZACIÓN EN INTELIGENCIA ARTIFICIAL (*AI LITERACY*)



Fuente: AILit framework, https://ailiteracyframework.org/es/

1. *Computación*: El fundamento de la IA es computacional;por tanto, los estudiantes deben aprender informática, pensamiento algorítmico, codificación, abstracción, etc.

2. *Comunicación digital*: Se refiere a la capacidad de acceder, analizar, evaluar, crear y comprender mensajes en los medios digitales de forma crítica, ya que la IA se utiliza cada vez más para crear contenido en redes sociales y otras plataformas de comunicación.

3. *Responsabilidad*: Relacionada con la seguridad y privacidad de los datos, la propiedad intelectual, los derechos y deberes, etc.

4. *Datos*: La IA moderna está basada en datos, se alimenta de ellos y los produce como respuesta. Por tanto, es necesario aprender sobre procesado de datos, sesgos, inferencias, modelos, probabilidad, etc.

5. *Diseño de soluciones*: Relacionado con la resolución de problemas mediante un ordenador, idealización, descomposición, iteración, etc.

6. *Ética*: Centrada en conocer el impacto de la IA, sus riesgos y beneficios sobre los humanos.

Personalmente, considero que falta una séptima área básica en el esquema anterior, ya que el enfoque *STEM* (Ciencia, Tecnología, Ingeniería y Matemáticas) debe formar también parte de las bases de la IA. Es decir, la formación en IA no debería estar sesgada por la tendencia actual hacia el uso de IA generativa, herramientas como ChatGPT y similares que "habitan" un entorno digital puro. En breve, la IA estará presente en el entorno físico, en robots domésticos, vehículos autónomos o casas inteligentes. Todo esto implica formar a los estudiantes en los principios básicos de la ingeniería, es decir, lograr que las soluciones tecnológicas resuelvan problemas en el mundo real. A nivel educativo, estamos hablando de un enfoque práctico, manipulativo, que implique el desarrollo de proyectos aplicados, como programar robots, crear aplicaciones para *smartphone* o automatizar una habitación (Bellas *et al*., 2023).

Comprender la IA, como comentaré más adelante, va más allá de comprender la IA generativa o el aprendizaje automático. Y, sobre todo, va más allá de las tendencias que marcan las grandes empresas tecnológicas. En educación no podemos ir a remolque de decisiones comerciales, pero tampoco podemos estar ajenos al avance. Por eso se requiere un planteamiento riguroso y a largo plazo, pero también flexible (Bellas, 2024).

2.3. ¿Cómo se debe enseñar?

La alfabetización en IA se puede articular, de forma general, con dos enfoques básicos. Por un lado, como usuarios de herramientas de IA. Los estudiantes pueden aprender a usar estas herramientas con gran destreza, sacándoles mucho provecho y comprendiendo sus resultados al adquirir el hábito del pensamiento crítico. Este es un enfoque que permite abarcar un contenido amplio, pero que depende del rumbo que lleve la tecnología, y que implica una adaptación constante.

Por otro lado, podrían aprender como creadores, es decir, programando sistemas simples de IA. Este segundo enfoque implica adquirir conocimientos sobre programación, ciencia de datos y STEM. Es decir, conlleva una mayor profundidad técnica, un avance menor en los contenidos, pero dota a los estudiantes de una base más sólida, necesaria a la hora de resolver problemas complejos y fundamental para afrontar los continuos cambios de esta disciplina.

Las recomendaciones actuales como (European Commission y OECD, 2025) o (Cosgrove y Cachia, 2025) dejan el segundo enfoque para cursos superiores, estudiantes avanzados o grupos especiales. Pero debemos recordar que este es el enfoque que se sigue con materias clásicas como las lenguas, las matemáticas o el arte, ¿por qué no la IA?[2]

2.4. ¿A quién enseñar?

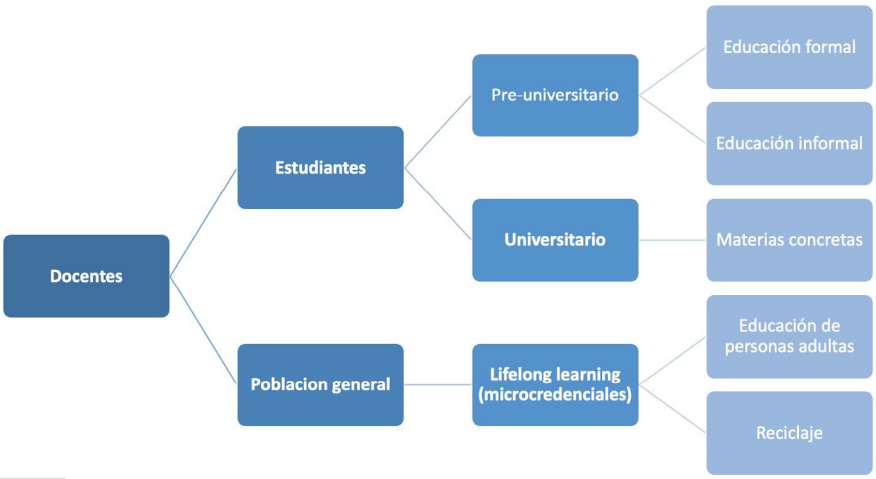
Idealmente, una alfabetización en IA debería llegar a todos los segmentos de la población, como se ilustra en la figura 3. Desde los estudiantes preuniversitarios en el sistema educativo reglado, pasando por los universitarios y los profesionales que desean reciclarse y tener nuevas oportunidades laborales, hasta llegar a la población adulta. Con independencia de las formaciones particulares que puedan dar las empresas privadas a sus empleados, a nivel público se deberían dar pasos para lograr esta alfabetización global si deseamos democratizar el uso de esta tecnología.

El primer paso es, obviamente, formar a los formadores. No existe aún una regulación cerrada y formal que marque una determinada formación en IA para todo el profesorado de secundaria o universidad, pero sí existe un marco competencial nacional (MRCDD) que incorpora IA (INTEF, 2022), unas orientaciones oficiales de INTEF (2025) que definen principios, objetivos y líneas de actuación para integrar la IA en la formación docente (INTEF, 2025), y un marco de referencia universitario impulsado por la CRUE (Conferencia de Rectores y Rectoras de las Universidades Españolas) y desarrollado mediante planes de formación propios de cada universidad (Crue Universidades Españolas, 2024).

[2] Para profundizar en este tema, recomiendo el trabajo de Jesús Moreno León, por ejemplo en https://www.linkedin.com/pulse/analfabetismo-en-tiempos-de-inteligencia-artificial-jes%C3%BAs-moreno-le%C3%B3n-gdmyf/

FIGURA 3

ESQUEMA DEL ALCANCE DE LA ALFABETIZACIÓN EN IA



Fuente: Elaboración propia.

2.5. ¿Cuándo enseñar?

En cuanto al cuándo alfabetizar en IA, el propio concepto implica planes educativos a largo plazo, por lo que la respuesta más general sería: desde la educación primaria hasta la universidad. La formación en IA debería ser un proceso incremental que cubriese todas las áreas comentadas antes, adecuando los contenidos al desarrollo cognitivo y la edad. Es con este enfoque con el que coincide la mayor parte de los sistemas educativos europeos, que buscan integrar los fundamentos de la IA en diversas materias para conseguir un avance gradual interdisciplinar (European Commission y OECD, 2025).

Pero la complejidad de articular esta alfabetización integrada y a largo plazo reside en la formación de los docentes, muchos de ellos no especialistas en tecnología. Por este motivo, algunos responsables educativos han optado por implementar ya materias concretas sobre IA, a modo de alfabetización *express*, que cubran los principales aspectos con docentes específicos. En el caso de España, la LOMLOE y el RD 217/2022 no incluyen IA como materia específica nacional, pero sí la materia obligatoria "Tecnología y Digitalización" en la ESO, que incorpora contenidos transversales de IA, algoritmos y pensamiento computacional. A nivel regional, Galicia es pionera con una materia específica de IA en 4.º de la ESO y otra en 1.º de bachiller desde 2023 (Xunta de Galicia, 2025; Xunta de Galicia, 2023), y Valencia incluye contenidos concretos en la materia de "Programación, Inteligencia Artificial y Robótica" (2.º, 3.º y 4.º de la ESO). Madrid y Andalucía están implantando o tienen en marcha materias similares. Todas son optativas y dependen de la oferta de cada centro educativo.

Actualmente, no existen iniciativas regladas en España entre educación primaria, ni universitaria, y mucho menos, en cuanto a educación de personas adultas.

3. CAPACITACIÓN, RESPONSABILIDAD Y CONTROL

3.1. Comprender qué es y qué no es

La necesidad de una alfabetización general en IA surge, en primer lugar, de la dificultad generalizada para comprender con rigor qué es y qué no es la IA. Para una parte muy significativa de la población, incluidos estudiantes y docentes, la IA es, fundamentalmente, ChatGPT. Aunque es algo totalmente comprensible, está muy alejado de la realidad del campo. De hecho, la definición de sistema de IA que encontramos en la Ley de IA es esta: "un sistema basado en máquinas que está diseñado para funcionar con diversos niveles de autonomía y que puede mostrar capacidad de adaptación tras su despliegue, y que, para objetivos explícitos o implícitos, infiere, a partir de la entrada que recibe, cómo generar salidas tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales"[3].

¿Cuántos ciudadanos comprenden realmente esta definición? ¿Cuántos la asociarían con ChatGPT? Los conceptos clave de la IA como autonomía, adaptación, predicción o entorno que se mencionan arriba son básicos para entenderla en toda su dimensión, más allá de las tendencias actuales y, sobre todo, del "ruido" de los medios generalistas. Aunque este enfoque global de la IA pueda sonar muy académico y alejado de la realidad de los usuarios, no es así. La historia de la IA[4], aunque joven, no comenzó en el año 2022, y continuará más allá de la IA generativa, por lo que debemos fomentar una alfabetización general y rigurosa.

Los principales textos sobre IA utilizan el enfoque de agente inteligente para proporcionar una base sólida y general del campo (Poole y Mackworth, 2023; Russell y Norvig, 2021), es decir, se plantea explicar la IA como sistema computacional autónomo situado en un entorno. Bajo este enfoque, un sistema de IA no se define por su apariencia externa ni por su complejidad aparente, sino por su capacidad para percibir información de dicho entorno, razonar o representar internamente esta información y actuar de manera adaptativa en función de objetivos explícitos o implícitos. Este marco permite explicar de forma coherente tanto los sistemas actuales como otros más complejos que puedan desarrollarse en el futuro, facilitando una comprensión escalable de la tecnología. Desde un vehículo autónomo hasta un sistema de diagnóstico médico conversacional, todos pueden analizarse bajo este mismo esquema, lo que reduce la dependencia de ejemplos aislados y favorece una comprensión transferible (Bellas *et al*., 2023).

[3] https://artificialintelligenceact.eu/es/article/3/

[4] https://es.wikipedia.org/wiki/Historia_de_la_inteligencia_artificial

Esta clarificación conceptual no tiene únicamente un valor técnico, sino que constituye la base para el desarrollo del pensamiento crítico frente a la IA. Saber distinguir cuándo un sistema actúa de manera adaptativa y cuándo ejecuta simplemente instrucciones preprogramadas permite evaluar con mayor criterio su fiabilidad, sus limitaciones y los riesgos asociados a su uso. Una población alfabetizada tiene criterio propio, y podrá saber si, al comprar un aspirador autónomo con IA, realmente es autónomo y realmente usa IA. Este ejercicio, aparentemente simple, resulta clave para trasladar la reflexión sobre la IA a contextos cotidianos, como el uso de aplicaciones móviles, sistemas de recomendación o herramientas automatizadas de decisión.

Es tiempo de que comprendamos la IA más allá de la identificación reduccionista que nos dan los medios, la ciencia ficción o los nuevos "expertos", que genera expectativas, temores o confianzas infundadas. La alfabetización en IA no busca que toda la población se convierta en experta técnica, sino que adquiera la capacidad de formular preguntas informadas, reconocer cuándo un sistema merece confianza y cuándo requiere supervisión humana. Este es, en última instancia, el fundamento sobre el que puede construirse un uso responsable, seguro y socialmente sostenible de la IA.

3.2. Sacar el máximo rendimiento a las herramientas

Vivimos en la era de la IA generativa, con nuevas herramientas para generar contenido digital apareciendo cada día. Es necesario, como hemos dicho, que una alfabetización en IA incluya también capacitación específica en el uso de estas herramientas. Pero, para no depender de los avances particulares de las empresas desarrolladoras, debemos centrar la formación en las características comunes y básicas detrás de la tecnología.

Por ejemplo, debemos saber que la IA generativa está basada en grandes modelos de lenguaje (LLM), y es importante conocer cómo funcionan (Funcas, 2025), así como sus limitaciones innatas, como la inexactitud, los sesgos, o la tendencia a la complacencia (Johnson y Hyland-Wood, 2024). Estas limitaciones se pueden reducir si hacemos un buen uso de las opciones de configuración, del *prompt* y, sobre todo, si añadimos contexto a la petición. También debemos conocer qué diferencia un modelo instantáneo de un modelo que razona, para poder utilizar uno u otro en función del problema a resolver.

Para sacar el máximo rendimiento a las herramientas actuales y futuras, planteo aquí la necesidad de una formación rigurosa[5]. Esto puede parecer evidente, pero no ha sido la práctica habitual en la educación digital, donde se ha fomentado el autoaprendizaje con resultados bastante mediocres a largo plazo.

[5] Por ejemplo, https://descargas.intef.es/cedec/proyectoedia/guias/contenidos/inteligencia_artificial/index.html

3.3. Tener pensamiento crítico

El desarrollo del pensamiento crítico constituye uno de los pilares fundamentales de cualquier proceso de alfabetización en IA. En un contexto mediático caracterizado por titulares llamativos y, en ocasiones, abiertamente alarmistas o excesivamente optimistas, la ciudadanía se enfrenta a narrativas contradictorias sobre el impacto de la IA. Algunas de estas narrativas anuncian transformaciones disruptivas inmediatas, mientras que otras prometen mejoras generalizadas sin costes ni riesgos. La alfabetización en IA debe proporcionar las herramientas necesarias para interpretar estos mensajes con rigor, permitiendo distinguir entre afirmaciones respaldadas por evidencia empírica sólida y aquellas basadas en extrapolaciones, estudios metodológicamente débiles o intereses comerciales.

Esta necesidad se hace particularmente evidente al analizar la literatura científica sobre el uso de la IA generativa en ámbitos como la educación. A pesar de la abundancia de publicaciones que anuncian beneficios significativos, la realidad es que, a día de hoy, no existe un cuerpo robusto de evidencia científica que demuestre de forma concluyente impactos positivos generalizados del uso de IA generativa en los procesos de aprendizaje. Muchos de los estudios disponibles presentan limitaciones importantes, como muestras reducidas, ausencia de grupos de control o evaluaciones centradas en percepciones subjetivas más que en resultados medibles. El pensamiento crítico implica, en este sentido, la capacidad de analizar con cautela estos trabajos, comprender sus alcances reales y evitar conclusiones precipitadas basadas en evidencias parciales (Holmes *et al*., 2025).

En el caso específico de la IA generativa, el pensamiento crítico adquiere una dimensión adicional: la evaluación de la calidad, fiabilidad y adecuación de las respuestas producidas por las herramientas de IA. Dado que estos modelos generan contenidos plausibles pero no necesariamente correctos, los usuarios deben ser capaces de juzgar el resultado en función de su coherencia, su alineación con el contexto y su correspondencia con conocimientos previos contrastados. Este juicio no es independiente del nivel de competencia del propio usuario: cuanto mayor es su dominio del ámbito en cuestión, mayor es su capacidad para detectar errores, omisiones o sesgos en la respuesta generada. Por ello, la alfabetización en IA debe enfatizar que estas herramientas no sustituyen al conocimiento, sino que lo amplifican o lo distorsionan en función del grado de control humano ejercido.

Finalmente, comprender las oportunidades y limitaciones de la IA generativa implica reconocer que muchas de sus carencias pueden mitigarse mediante un uso más competente y consciente de la herramienta. La formulación adecuada de preguntas, la verificación cruzada de resultados y la integración de la IA como apoyo y no como sustituto del razonamiento humano son habilidades que se desarrollan con formación y práctica, es decir, con capacitación técnica. Desde esta perspectiva, el pensamiento crítico no es

un complemento opcional de la alfabetización en IA, sino su núcleo central: es lo que permite que la tecnología se convierta en un instrumento de mejora y no en una fuente de dependencia o desinformación.

3.4. Conocer derechos y responsabilidades

La alfabetización en IA no puede considerarse completa sin una formación básica en los aspectos legales que regulan su diseño, despliegue y uso. En el contexto europeo, la aprobación de la Ley de IA introduce un marco normativo que afecta de manera directa no solo a los desarrolladores de sistemas, sino también a las personas y organizaciones que los utilizan en contextos profesionales. Conocer la existencia de esta regulación, sus principios generales y sus implicaciones prácticas resulta imprescindible para que la ciudadanía y, en particular, los profesionales de sectores sensibles como la educación, la banca o la administración pública, puedan ejercer un uso responsable de la tecnología y evitar riesgos legales, éticos y organizativos (Agencia Española de Supervisión de Inteligencia Artificial —AESIA—, 2026).

Uno de los elementos más relevantes del marco normativo europeo es la introducción del concepto de *deployer*, es decir, la figura responsable del uso efectivo de un sistema de inteligencia artificial en un contexto determinado. Esta noción rompe con la idea, aún muy extendida, de que la responsabilidad recae exclusivamente en quien diseña o comercializa la tecnología. En ámbitos como el educativo, por ejemplo, los docentes que emplean herramientas de IA con sus estudiantes son *deployers* en este contexto, y los equipos directivos que ponen estas herramientas a disposición de los docentes, también (Roddeck y Ooge, 2025). Esto implica obligaciones concretas, como informar sobre el uso de la IA, justificar su idoneidad y garantizar que sus resultados no vulneren derechos fundamentales ni principios de equidad. La sorpresa que este hecho genera en muchos profesionales pone de manifiesto la brecha existente entre la adopción tecnológica y el conocimiento del marco legal que la regula.

Desde esta perspectiva, la formación en los aspectos legales de la IA cumple una función preventiva y habilitadora. Preventiva, porque permite anticipar riesgos y evitar usos indebidos que podrían derivar en responsabilidades legales o daños reputacionales; y habilitadora, porque dota a los usuarios de criterios claros para integrar la IA de forma legítima y transparente en sus prácticas profesionales. Comprender que la Ley de IA no se dirige únicamente a grandes empresas tecnológicas, sino que establece un sistema de responsabilidad distribuida, es un paso esencial para fomentar una cultura de uso consciente y controlado. La alfabetización legal en IA no persigue formar juristas, sino capacitar a los usuarios para tomar decisiones informadas y alineadas con el marco normativo vigente, reforzando así la confianza social en el uso de estas tecnologías.

4. EL IMPACTO DE LA IA EN EL APRENDIZAJE

El avance en el uso práctico de la IA generativa está mostrando que, cuando una persona domina previamente un ámbito de conocimiento y utiliza adecuadamente herramientas basadas en esta tecnología como apoyo, el resultado no es únicamente una mayor eficiencia, sino una mejora cualitativa del trabajo realizado. Al automatizar tareas repetitivas, tediosas o de bajo valor añadido para el humano, la IA permite concentrar el esfuerzo en decisiones de mayor nivel, potenciando así la calidad del resultado final. Este efecto potenciador, sin embargo, no es universal ni automático, sino que depende de la relación entre la herramienta y el nivel de competencia del usuario. La clave reside en que el control del proceso sigue siendo humano: la persona decide qué aceptar, qué descartar y cómo integrar la aportación de la IA en un contexto más amplio de conocimiento y experiencia (New Cartographies, 2025).

No obstante, este equilibrio se rompe cuando la IA comienza a asumir funciones centrales en tareas que requieren práctica continuada para mantener la competencia. En situaciones en las que una habilidad necesita ejercitarse de forma regular, como ocurre en determinados procedimientos técnicos o en la resolución de problemas complejos, el uso sistemático de sistemas automatizados puede conducir a una pérdida progresiva de destreza. La analogía con otros ámbitos, como la aviación o la cirugía asistida por robots, ilustra bien este riesgo: cuando una parte sustancial del proceso se delega de forma constante en la tecnología, la capacidad humana para intervenir de manera autónoma se debilita, y deben ser incluidos periodos de práctica sin la tecnología para no perder la habilidad adquirida.

En aquellas situaciones donde la IA deja de ser una herramienta de apoyo al trabajo y pasa a convertirse en un sustituto funcional, entonces el humano pierde toda relevancia. Sin la tecnología, el trabajo no se hace. La alfabetización en IA tiene aquí poco impacto, al no existir un ciclo de interacción entre humanos y máquinas.

En el caso del aprendizaje, muchos estudiantes creen que, al utilizar IA para hacer un trabajo académico, están haciendo un uso similar al que hace un experto y que comentamos arriba. Es decir, creen que dominan el tema y la IA les potencia su aprendizaje, pero la realidad es que se encuentran en supuesto final, donde la IA hace todo el trabajo. Aprender es un proceso fisiológico bien estudiado (Ruiz Martín, 2020a), donde la IA tiene el potencial de aportar beneficios, pero no lo puede acelerar. Utilizar estas herramientas para proporcionar la respuesta en un trabajo académico sin llevar a cabo el proceso por uno mismo no contribuye en absoluto al aprendizaje.

Comprender la diferencia entre usar la IA como apoyo al aprendizaje y usarla como sustituto del aprendizaje es esencial para preservar la autonomía intelectual y garantizar que esta contribuya realmente a una mejora sostenible de los procesos educativos. Muchos docentes pueden pensar que este uso "positivo" de la IA es utópico, ya que los estudian-

tes la están utilizando justo para lo contrario, el plagio, atribuyendo a un tercero (la herramienta) la autoría de un trabajo propio. Pero el problema aquí no es solo de la integridad académica de los estudiantes, sino de la falta de reflexión de muchos docentes.

En diferentes niveles educativos, mayormente educación secundaria y universitaria, ha sido práctica habitual las tareas evaluables consistentes en la redacción de un texto o informe para ser realizadas de manera autónoma fuera del aula. Estas tareas pueden implicar búsqueda en fuentes digitales (internet), recopilación de ideas, resumen y redacción/presentación. Con diferentes niveles de utilidad, han sido metodologías docentes habituales en materias generalistas como historia, ciencias o lengua, y en otras más específicas en cuanto a la educación superior. La aparición de la IA generativa y su uso tan extendido entre los jóvenes ha anulado completamente la utilidad de este tipo de actividad educativa, ya que son muchos los estudiantes que utilizan herramientas como ChatGPT y similares para su realización. Como he explicado en el punto anterior, esto anula totalmente el interés del trabajo, ya que no hay aprendizaje en este proceso (podríamos debatir este punto si el objetivo de la tarea está en aprender a usar la IA generativa, claro). Es, por tanto, responsabilidad del docente cambiar este tipo de metodología, por otra que se realice en el aula sin acceso a la IA, o bien por otra que la integre de alguna forma que no impacte en el aprendizaje. Lo que es obvio es que no podemos prohibir el uso de la IA fuera del aula, por tanto, debemos analizar su impacto y obrar en consecuencia.

¿Para qué enseñar sobre IA generativa, entonces? Si no vamos a permitir su uso en el aula, ¿qué interés tiene que los estudiantes la aprendan a utilizar? Pues precisamente es en el escenario que se propone en este artículo donde gana interés: una vez que no tiene utilidad para la realización de las tareas autónomas, podemos enseñar a los estudiantes a usarla para potenciar el aprendizaje. Volvemos pues, a la idea de usar la IA como asistente y aliada, no como sustituta.

Para ilustrar esta idea de una forma concreta, voy a utilizar el trabajo en ciencias cognitivas aplicadas al aprendizaje como base (Ruiz Martín, 2020b). Según la evidencia científica existente, un aprendizaje profundo se caracteriza por ser duradero, transferible, funcional y productivo, es decir, perdura en la memoria, se aplica en distintos contextos, permite hacer cosas nuevas con él y facilita futuros aprendizajes. Para lograrlo, la investigación en ciencias cognitivas destaca estrategias como la evocación (recuperar activamente información de la memoria), la elaboración (reflexionar sobre el significado y conectar las ideas nuevas con conocimientos previos) y la transferencia (identificar o emplear las mismas ideas o procedimientos en contextos diversos) (Ruiz Martín, 2020a). Veamos un ejemplo sencillo de uso de un *chatbot* de IA generativa para ayudar a un estudiante de bachillerato a preparar diferentes exámenes utilizando estas estrategias:

- Evocación: "Te voy a subir mis apuntes sobre fotosíntesis. Sin añadir información externa, genera 5 preguntas cortas que apunten a lo que mi profesora ha destacado para el examen (cloroplasto, ATP/NADPH, ciclo de Calvin, estomas,

factores limitantes). Espera a que yo responda. Después, corrige comparando mis respuestas con los apuntes: indica correcto/incorrecto/parcial, añade la frase o idea exacta que lo justifica y, si algo no aparece en los apuntes, escribe literalmente 'No consta en los apuntes'."

- Elaboración: "La semana que viene tengo un examen sobre pronombres. Propón una situación comunicativa (por ejemplo, una queja formal y una conversación breve) y pídeme que escriba un texto de 120–150 palabras donde tenga que usar, al menos una vez y de forma natural, pronombres personales tónicos y átonos, demostrativos, posesivos, relativos e indefinidos (y, si procede, 'se'). Después: subraya y clasifica cada pronombre, explica su función sintáctica, señala dos errores frecuentes y cómo evitarlos."

- Transferencia: "Para mi examen de Educación Física sobre el calentamiento previo al ejercicio, te paso mis materiales de clase. Con esa base, plantea un caso realista (deporte, objetivo de la sesión, duración, nivel del grupo, espacio/material y una limitación: por ejemplo, hace frío o hay un alumno con molestias de tobillo). Yo diseñaré un calentamiento completo (general + específico) con tiempos, progresión e intensidad. Después, revísalo: comprueba coherencia con el objetivo, seguridad, progresión, activación de cadenas musculares implicadas y transferencia al gesto deportivo; propón mejoras y justifícalas."

Si pensamos que es una redacción muy compleja para un estudiante de esta edad, es que no hemos comprendido aún qué significa la capacitación en el uso de estas herramientas: todos estos *prompts* han sido generados por el propio *chatbot*, pidiéndole maximizar la utilidad de la respuesta. Es decir, el estudiante futuro, alfabetizado adecuadamente en IA, sabrá cómo le puede ayudar a aprender (evocar, elaborar, transferir), lo relevante que es añadir contexto y limitar la creatividad (sus apuntes), y que la mejor forma de obtener una respuesta útil es haciendo que el *prompt* final lo proporcione la propia IA.

5. CONCLUSIONES

La rápida expansión de la inteligencia artificial ha situado a la sociedad ante una transformación tecnológica de alcance estructural, cuyos efectos se manifiestan ya en el ámbito laboral, educativo y social. Frente a un modelo de adopción impulsado principalmente por incentivos económicos y por la competitividad entre grandes actores tecnológicos, este artículo plantea la alfabetización en IA como un mecanismo imprescindible de equilibrio, capaz de introducir control, responsabilidad y criterio humano en el uso de estas tecnologías.

A lo largo del texto se argumenta que la alfabetización en IA no puede limitarse a la capacitación instrumental en herramientas concretas, ni centrarse exclusivamente en las tendencias actuales, como la IA generativa. Por el contrario, debe apoyarse en una comprensión sólida de los fundamentos conceptuales de la IA, en el desarrollo del pensamiento crítico y en el conocimiento de los derechos y responsabilidades que introducen los marcos legales.

Solo desde esta base es posible que los usuarios (estudiantes, profesionales y ciudadanía en general) evalúen con rigor tanto los beneficios como los riesgos de la IA, evitando posiciones acríticas o reacciones alarmistas. Especial atención merece el impacto de la IA en los procesos de aprendizaje y en la autonomía intelectual. El análisis realizado muestra que la IA puede actuar como un potente amplificador del conocimiento experto cuando existe una competencia previa sólida, pero también como un sustituto del aprendizaje cuando se utiliza de forma prematura o acrítica. Esta distinción resulta clave en el ámbito educativo, donde el uso indiscriminado de herramientas de IA puede comprometer el desarrollo de habilidades fundamentales si no se integra dentro de un enfoque pedagógico consciente y gradual.

Finalmente, la alfabetización en IA debe entenderse como una inversión social a largo plazo. No se trata de formar especialistas en tecnología, sino de capacitar a la población para convivir con sistemas inteligentes de forma informada, crítica y responsable. En este sentido, la educación se confirma, una vez más, como el principal instrumento para garantizar que el avance tecnológico redunde en mayor autonomía, equidad y bienestar social, y no en dependencia, opacidad o pérdida de control humano.

Referencias

Agencia Española de Supervisión de Inteligencia Artificial —AESIA —. (2025). *Garantizando una IA ética y responsable*. Consultado el 5 de enero de 2026. https://aesia.digital.gob.es/es

Agencia Española de Supervisión de Inteligencia Artificial —AESIA—. (2026). *Guías*. Consultado el 5 de enero de 2026. https://aesia.digital.gob.es/es/guias

AI Singapore. (2025). *Learn AI Singapore: Comprehensive Platform for AI Literacy*. Consultado el 5 de enero de 2026. https://learn.aisingapore.org

Appio, F. P., La Torre, D., Lazzeri, F., Masri, H., & Schiavone, F. (Eds.). (2023). *Impact of Artificial Intelligence in Business and Society: Opportunities and Challenges* (1st ed.). Routledge. https://doi.org/10.4324/9781003304616

Australian Education Council. (2025). *Australian Framework for Generative Artificial Intelligence in Schools*. Revised and endorsed in June 2025. Consultado el 5 de enero de 2026. https://www.education.gov.au/schooling/resources/australian-framework-generative-artificial-intelligence-ai-schools

Bellas, F., Guerreiro-Santalla, S., Naya, M., *et al.* (2023). AI Curriculum for European High Schools: An Embedded Intelligence Approach. *International Journal of Artificial Intelligence in Education*, 33, 399–426. https://doi.org/10.1007/s40593-022-00315-0

Bellas, F. J. (2024). Educar para la inteligencia artificial: un enfoque en perspectiva. En Ò. Flores-Alarcia, y L. Fornons Casol (eds.), *Educación e Inteligencia Artificial: Horizontes de transformación*. Madrid: Dykinson. https://www.dykinson.com/libros/educacion-e-inteligencia-artificial-horizontes-de-transformacion/9788410708778/

Bellas, F. J. (2025). Explainable artificial intelligence and its key role in education: Promoting critical thinking and autonomy in the classroom. *Metode Science Studies Journal, 15*(7), e30514. https://doi.org/10.7203/metode.15.30514

Comisión Europea, Agencia Ejecutiva Europea de Educación y Cultura. (2025). *Explainable AI in education: Fostering human oversight and shared responsibility: by the European Digital Education Hub's Squad on artificial intelligence in education*. Oficina de Publicaciones de la Unión Europea. https://data.europa.eu/doi/10.2797/6780469

Corredera, J. C. (2023). Inteligencia artificial generativa. *Anales de la Real Academia de Doctores de España, 8*(3), 475–489. Consultado el 5 de enero de 2026. https://www.rade.es/imageslib/PUBLICACIONES/ARTICULOS/V8N3

Cosgrove, J., & Cachia, R. (2025). *DigComp 3.0: European Digital Competence Framework – Fifth edition*. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0001149

Crue Universidades Españolas. (2024). *Digitalización e Inteligencia Artificial Generativa*. Crue Universidades Españolas. https://www.crue.org/wp-content/uploads/2024/03/Crue-Digitalizacion_IA-Generativa.pdf.

ENAE International Business School. (2025). *La inteligencia artificial en la vida cotidiana*. Blog de ENAE International Business School. Consultado el 5 de enero de 2026. https://www.enae.es/blog/la-inteligencia-artificial-en-la-vida-cotidiana

European Commission y OECD. (2025a). AI Literacy Framework for Primary and Secondary Education. *Documento de trabajo*. Publications Office of the European Union. Consultado el 5 de enero de 2026. https://joint-research-centre.ec.europa.eu/ai-literacy-framework_en

European Commission and OECD. (2025b). Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education. *Documento de trabajo.* Publications Office of the European Union. Consultado el 5 de enero de 2026. https://ailiteracyframework.org

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. Consultado el 5 de enero de 2026. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

European Union. (2025). Regulation (EU) 2024/1689: Article 4 - AI Literacy. AI Act provisions on AI literacy came into force in February 2025. Consultado el 5 de enero de 2026. https://digital-strategy.ec.europa.eu/en/policies/ai-talent-skills-and-literacy

Funcas. (2025). *Grandes modelos de lenguaje: de la predicción de palabras a la comprensión*. Funcas. Consultado el 5 de enero de 2026. https://www.funcas.es/articulos/grandes-modelos-de-lenguaje-de-la-prediccion-de-palabras-a-la-compresion/

Holmes, W., Mouta, A., Hillman, V., Schiff, D., Laak, K.-J., Atenas, J., Bardone, E., Lochead, K., Gonsales, P., Havemann, L., Seon, J., Go, B., Schreurs, B., Zhgenti, S., Lee, K., Bali, M., Bialik, M., Medina-Gual, L., Knight, S., *et al.* (2025). *Critical Studies of Artificial Intelligence and Education: Putting a Stake in the Ground*. The Education Research Network SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5391793

Innovation, Science and Economic Development Canada. (2025). *Learning Together for Responsible Artificial Intelligence*. National working group on AI literacy and responsible AI. Consultado el 5 de enero de 2026. https://ised-isde.canada.ca/site/advisory-council-artificial-intelligence/en/public-awareness-working-group/learning-together-re

INTEF. (2022). *Marco de Referencia de la Competencia Digital Docente*. Instituto Nacional de Tecnologías Educativas y de Formación del Profesorado (INTEF). Consultado el 5 de enero de 2026. https://intef.es/Noticias/marco-de-referencia-de-la-competencia-digital-docente/

INTEF. (2025). *Orientaciones para la integración de la inteligencia artificial en la formación docente*. Instituto Nacional de Tecnologías Educativas y de Formación del Profesorado (INTEF). https://intef.es/wp-content/uploads/2025/06/Orientaciones-para-la-integracion-de-la-IA-en-la-formacion-docente-2-1.pdf

Johnson, S., & Hyland-Wood, D. (2024). *A Primer on Large Language Models and their Limitations*. arXiv preprint arXiv:2412.04503. https://arxiv.org/abs/2412.04503

Microsoft and Brazilian Government. (2025). *ConectAI: AI Literacy Training Initiative in Brazil*. Program targeting 5 million learners by 2027. Consultado el 5 de enero de 2026. https://news.microsoft.com/source/latam/features/ai/microsoft-ai-courses-brazil/

Ministry of Education of China. (2025). *General AI Education Guide for Primary and Secondary Schools*. Guidance issued in May 2025 for mandatory AI education

implementation. Consultado el 5 de enero de 2026. https://english.www.gov.cn/policies/policywatch/202504/18/content_WS6801bda9c6d0868f4e8f1da9.html

New Cartographies. (2025). *The Myth of Automated Learning*. Consultado el 6 de enero de 2026. https://www.newcartographies.com/p/the-myth-of-automated-learning

OECD. (2025). *PISA 2029 Media and Artificial Intelligence Literacy (MAIL) Assessment*. International student assessment program. Consultado el 5 de enero de 2026. https://www.oecd.org/en/about/projects/pisa-2029-media-and-artificial-intelligence-literacy.html

OpenAI. (2022). Presentamos ChatGPT. Blog post and research release, November 29, 2022. Consultado el 5 de enero de 2026. https://openai.com/es-ES/index/chatgpt/

Poole, D. L., & Mackworth, A. K. (2023). *Artificial Intelligence: Foundations of Computational Agents* (3rd ed.). Cambridge University Press.

ProFuturo and OEI. (2025). *AI in Latin American Classrooms: Mapping a Future Under Construction*. Regional analysis of 26 AI literacy initiatives. Consultado el 5 de enero de 2026. https://profuturo.education/en/observatory/inspiring-experiences/ia-en-las-aulas-latinoamericanas-mapeando-un-futuro-en-construccion

Roddeck, L., Bellas, F., Abu-Rasheed, H., & Ooge, J. (2025). Educational Stakeholders Need a Legal Guide on Explainable AI. Trabajo presentado en el Workshop *XAI-Ed: Pedagogy-Founded Explainable AI for Transparent, User-Centred AI in Education*, *26th International Conference on Artificial Intelligence in Education (AIED), Palermo, Italia, julio de 2025.* Consultado el 5 de enero de 2026. https://www.researchgate.net/publication/395459156_Educational_Stakeholders_Need_A_Legal_Guide_on_Explainable_AI

Ruiz Martín, H. (2020a). *¿Cómo aprendemos? Una aproximación científica al aprendizaje y la enseñanza*. Graó (Educación basada en evidencias, Serie Fundamentos de la Educación). ISBN: 978-84-18058-05-9.

Ruiz Martín, H. (2020b). *Conoce tu cerebro para aprender a aprender: Guía para jóvenes estudiantes*. Learning Bits S.L. ISBN: 978-84-122134-0-9.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. (2025). *The state of AI: Global survey 2025*. McKinsey Company (QuantumBlack, AI by McKinsey). Consultado el 5 de enero de 2026. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

StreamLex. (2025). *Insights from AI Literacy Practices Implemented by 15 Organizations*. European initiatives repository under EU AI Act Article 4. Consultado el 5 de enero de 2026. https://streamlex.eu/news/insights-from-ai-literacy-practices-implemented-by-15-organizations/

UNESCO. (2019). *Beijing Consensus on Artificial Intelligence and Education*. UNESCO Publishing. Consultado el 5 de enero de 2026. https://unesdoc.unesco.org/ark:/48223/pf0000368303

UNESCO. (2021). *Recomendación sobre la Ética de la Inteligencia Artificial*. United Nations Educational, Scientific and Cultural Organization (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa.

UNESCO. (2024). *AI Competency Frameworks for Teachers and Students*. UNESCO Publishing. Consultado el 5 de enero de 2026. https://www.unesco.org/en/articles/what-you-need-know-about-unescos-new-ai-competency-frameworks-students-and-teachers

The White House. (2025). Executive Order on Advancing Artificial Intelligence Education for American Youth. Presidential Action, April 23, 2025. Consultado el 5 de enero de 2026. https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth/

World Economic Forum. (2025). *The Future of Jobs Report 2025*. World Economic Forum. Consultado el 5 de enero de 2026. https://www.weforum.org/reports/the-future-of-jobs-report-2025

Xunta de Galicia (2025). *La Xunta refuerza los currículos de ESO y Bachillerato con tres nuevas materias optativas*. Nota de prensa oficial. Consultado el 5 de enero de 2026. https://www.xunta.gal/es/notas-de-prensa/-/nova/79579/xunta-refuerza-los-curriculos-eso-bachillerato-con-tres-nuevas-materias-optativas

Xunta de Galicia, Consellería de Cultura, Educación e Universidade. (2023). *Currículo oficial "IA para la Sociedad"*. Xunta de Galicia. Consultado el 4 de enero de 2026. https://www.edu.xunta.gal/portal/sites/web/files/content_type/learningobject/2023/09/08/f58cde00791509b31276b36b7e233213.pdf

CAPÍTULO IV

Agentic AI y Multiagentic: ¿Estamos reinventando la rueda? Una reinterpretación epistemológica

Vicent Botti

Los términos "Agentic AI (IA agentiva)" y "Multiagentic AI (IA multiagentiva)" han ganado popularidad recientemente en los debates sobre inteligencia artificial generativa, a menudo utilizados para describir agentes de *software* autónomos y sistemas compuestos por dichos agentes. Sin embargo, estos conceptos no representan una nueva ontología de agentes, sino una reconfiguración tecnológica de ideas ya consolidadas en la teoría de agentes autónomos y sistemas multiagente (MAS) desde hace muchos años. El uso de estos términos confunde estos conceptos de moda con conceptos bien establecidos en la literatura de la IA: agentes autónomos y sistemas multiagente. El capítulo aboga por el rigor científico y tecnológico y por el uso de la terminología establecida del estado del arte en IA, incorporando la riqueza del conocimiento existente —incluidos los estándares para plataformas de sistemas multiagente, lenguajes de comunicación y algoritmos de coordinación/cooperación, tecnologías de acuerdo (negociación automática, argumentación, organizaciones virtuales, confianza, reputación, etc.)— en la nueva y prometedora ola de agentes de IA basados en LLM, para no terminar reinventando la rueda.

Palabras clave: inteligencia artificial, agentes autónomos, sistemas multiagente, IA Generativa, LLMs.

1. INTRODUCCIÓN

El campo de los sistemas multiagente comenzó a tomar forma a principios de la década de 1990, inspirado por la idea de que el futuro de la inteligencia artificial implicaría redes de entidades de IA autónomas —conocidas como agentes— que actuarían de forma independiente para cumplir las tareas asignadas por los usuarios. Un elemento central de esta visión era la expectativa de que estos agentes necesitarían interactuar entre sí para alcanzar sus objetivos, lo que impulsó una investigación significativa para dotarlos de capacidades sociales como la cooperación, la coordinación, la negociación, la argumentación, la confianza y la reputación. Este capítulo ofrece un análisis crítico de este mal uso conceptual del término Agentic (agentiva) y defiende la invariancia conceptual del paradigma de agentes autónomos frente a la evolución tecnológica, examina la continuidad conceptual entre los agentes autónomos clásicos y las nuevas arquitecturas basadas en los LLMs, integrando las aportaciones epistemológicas y neurosimbólicas. Revisamos los orígenes teóricos de lo "agentic (agentiva)" en las ciencias sociales (Bandura, 1986) y las nociones filosóficas de intencionalidad (Dennett, 1971), y luego resumimos los trabajos fundamentales sobre agentes inteligentes y sistemas multiagente de Wooldridge, Jennings y otros. Examinamos las arquitecturas clásicas de agentes —desde simples agentes reactivos hasta modelos de Creencia-Deseo-Intención (BDI)— y destacamos las propiedades clave (autonomía, reactividad, proactividad, capacidad social) que definen la agencia en la IA. A continuación, discutimos los desarrollos recientes en los grandes modelos de lenguaje (LLM) y las plataformas de agentes basadas en LLM, incluida la aparición de agentes de IA basados en LLM y los marcos de orquestación multiagente de código abierto. Argumentamos que el término "Agentic AI (IA agentiva)" se utiliza a menudo como un término de moda para lo que son esencialmente agentes de IA, y "Multiagentic AI (IA multiagentiva)" para lo que son sistemas multiagente. Esta confusión pasa por alto décadas de investigación en el campo de los agentes autónomos/sistemas multiagente. Se sostiene que la IA agentiva no redefine la agencia, sino que la reinterpreta en clave cognitivo-lingüística, manteniendo la estructura funcional de percepción, decisión y acción de la abstracción de agente definida en el área de los agentes autónomos/sistemas multiagente. Finalmente, se discuten los riesgos de la inflación terminológica y las oportunidades de una síntesis neurosimbólica para el futuro de la inteligencia artificial. El capítulo aboga por el rigor científico y tecnológico y por el uso de la terminología establecida del estado del arte en IA, incorporando la riqueza del conocimiento existente —incluidos los estándares para plataformas de sistemas multiagente, lenguajes de comunicación y algoritmos de coordinación/cooperación, tecnologías de acuerdo (negociación automatizada, argumentación, organizaciones virtuales, confianza, reputación, etc.)— en la nueva y prometedora ola de agentes de IA basados en LLM, para no terminar reinventando la rueda.

Los recientes avances en la IA generativa, en particular los grandes modelos de lenguaje (LLM), han llevado a un resurgimiento del interés en los agentes de *software* autóno-

mos: sistemas de IA que pueden percibir de forma autónoma su entorno, razonar y actuar para cumplir sus objetivos de diseño y realizar tareas en nombre de los usuarios. Existe un creciente entusiasmo en torno al desarrollo de agentes basados en LLM que aprovechan su inteligencia general para operar de forma autónoma. Visionarios como Bill Gates predijeron que, en un futuro cercano, todos tendremos un asistente personal de IA "muy superior a la tecnología actual", capaz de responder a solicitudes en lenguaje natural y realizar diversas tareas comprendiendo los objetivos del usuario (Gates, 2023). Gates señala que este *software*, que él y otros han imaginado durante décadas, por fin se está convirtiendo en algo práctico gracias a los avances en IA, y que tales "agentes" podrían revolucionar nuestra interacción con los ordenadores. Además, hay una creciente evidencia de que organizar sistemas como conjuntos de agentes especializados e interactivos puede ser un enfoque eficaz para abordar tareas complejas de resolución de problemas. Paralelamente, cierto discurso industrial, no siempre (hay muchos casos en los que la terminología se respeta), ha introducido nuevos términos como "Agentic AI (IA agentiva)" para describir sistemas de IA dotados de autonomía y toma de decisiones proactiva. Blogs y artículos técnicos diferencian entre "agentes de IA" y "Agentic AI (IA agentiva)", presentando esta última como un marco general donde operan múltiples agentes. Los sistemas basados en los LLM, capaces de generar lenguaje y planificar acciones, han sido denominados "agentes agentivos" (una tautología que debilita la precisión científica), sugiriendo un salto cualitativo en autonomía y razonamiento. De manera similar, el término "Multiagentic (multiagentiva)" se utiliza a veces para sistemas con múltiples agentes que interactúan, de forma análoga a los sistemas multiagente clásicos.

Esta aparición de la terminología "agentic" en la IA plantea preguntas críticas. ¿Son estos conceptos genuinamente nuevos o simplemente nuevas etiquetas para ideas ya consolidadas en la investigación de agentes autónomos? El campo de los agentes inteligentes y los sistemas multiagente (MAS) tiene una rica historia que se remonta a décadas, con sus propios fundamentos teóricos, arquitecturas e incluso conferencias, revistas y asociaciones científicas dedicadas a investigadores en el área. Antes de abrazar el hype de la "Agentic AI", es importante examinar si el término se está aplicando erróneamente como una palabra de moda para capacidades comprendidas desde hace mucho tiempo en la IA. En otras palabras, ¿estamos reinventando la rueda al equiparar la Agentic IA con los agentes inteligentes y los sistemas multiagentic con los sistemas multiagente? La llamada IA agentiva constituye una continuidad conceptual del paradigma de agentes autónomos, aunque potenciada por nuevas tecnologías, la incorporación de la tecnología de los LLM como medio para que un agente determine, a partir de sus percepciones, qué acciones desempeñar para alcanzar los objetivos que le hayan sido delegados. Por tanto, la discusión actual no gira en torno a la definición del concepto de agente, sino a la reinterpretación de su implementación en contextos neurosimbólicos y lingüísticos.

2. LOS ORÍGENES DE AGENTIC: AGENCIA EN PSICOLOGÍA Y FILOSOFÍA

El adjetivo "agentic" ganó prominencia formal fuera de la informática. La primera referencia al término "agentic" aparece en el contexto de la psicología social, en el trabajo de psicólogos como Richard H. Goffman y Susan Fiske en la década de 1970, aunque sus raíces provienen de conceptos anteriores relacionados con la agencia y la autonomía. El término proviene de "agencia" y se refiere a la capacidad de los individuos para actuar de forma autónoma, tomar decisiones y ejercer control sobre sus vidas. El concepto de agencia ha existido durante mucho más tiempo, pero el término "agentic" como adjetivo se popularizó en la literatura académica especialmente a través del trabajo de Albert Bandura. En su libro *Social Foundations of Thought and Action: A Social Cognitive Theory* (1986), Albert Bandura introdujo el término agentic en el contexto de la agencia humana (Bandura, 1986). En la teoría social cognitiva de Bandura, agentic se refiere a la capacidad de los individuos para actuar intencionadamente, tomar decisiones y ejercer control sobre su entorno. Esto enfatiza un aspecto proactivo y deliberativo del comportamiento humano: en resumen, la cualidad de tener agencia. Una persona con capacidad agentic (agencia) no solo reacciona a fuerzas externas; son originadores de acciones, persiguiendo objetivos por su propia voluntad. La introducción del término por parte de Bandura destacó la naturaleza proactiva y autorreguladora de la acción humana, subrayando que las personas son agentes de su propio cambio. Este concepto de agencia ha sido influyente en la psicología, sustentando teorías de autoeficacia y motivación, pero originalmente se aplicó a humanos, no a máquinas.

Si buscamos en el diccionario, encontramos el siguiente significado de agentic:

- "Agentic" es un adjetivo derivado del sustantivo *agency*, y se usa comúnmente en psicología, sociología y educación. Se refiere a tener la capacidad de actuar de forma independiente, tomar decisiones y ejercer control sobre la propia vida y circunstancias.

- Definición (en contexto):

  - Que exhibe agencia; capaz de acción intencional y toma de decisiones.

  - Que actúa como un agente (es decir, iniciando y dirigiendo el propio comportamiento).

En paralelo, filósofos como Daniel C. Dennett exploraban cómo las atribuciones de agencia e intencionalidad podían aplicarse a máquinas y otros sistemas no humanos. El artículo de Dennett de 1971, *Sistemas Intencionales*, articuló la idea de que a menudo interpretamos y predecimos el comportamiento de sistemas complejos tratándolos como si tuvieran creencias, deseos e intenciones (Dennett, 1971). Dennett llamó sis-

tema intencional a aquel cuyo comportamiento puede explicarse y predecirse con éxito atribuyéndole cualidades mentalistas como creencias y deseos, de forma similar a la psicología popular que usamos para los humanos (Dennett, 1971). Esto proporciona una base filosófica para hablar de agentes de IA: incluso si un termostato o un programa no tienen literalmente deseos o creencias, tratarlo como un agente intencional (con objetivos, conocimiento, etc.) puede ser una abstracción útil para comprender su comportamiento (Dennett, 1971). La teoría de Dennett tendió así un puente entre la psicología humana y las entidades artificiales, sugiriendo que cualquier sistema cuyo comportamiento sea suficientemente complejo y dirigido puede ser visto en términos intencionales (Dennett, 1971).

Las nociones de Bandura (Bandura, 1986) y Dennett (Dennett, 1971) proporcionan un telón de fondo para la agencia como concepto: la agencia implica una acción intencional y dirigida a un objetivo, ya sea en humanos o en sistemas artificiales. Cuando llamamos a un sistema de IA agentivo, en el sentido estricto, queremos decir que exhibe agencia, es decir, que puede actuar de forma autónoma con la intención de alcanzar objetivos. En la psicología humana, esto es incontrovertido, pero en la IA plantea la pregunta: ¿qué se necesita para que un sistema artificial sea considerado un agente? El campo de la IA ha respondido a esto a lo largo de los años definiendo la abstracción del agente inteligente, que examinaremos a continuación. Baste decir que, para cuando el término agentic comenzó a aplicarse a la IA en la década de 2020, la comunidad ya tenía un profundo conocimiento de lo que significa que un sistema tenga agencia.

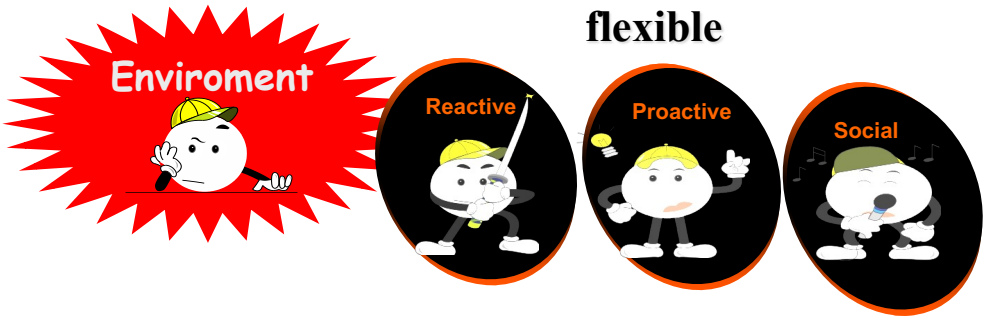
3. AGENTES INTELIGENTES: DEFINICIONES Y PROPIEDADES CLAVE

En 1991, Yoav Shoham propuso un nuevo paradigma de programación, la "Programación Orientada a Agentes" (Shoham, 1993). El paradigma promueve una visión social de la computación, en la que múltiples agentes interactúan entre sí.

En inteligencia artificial, un agente inteligente es fundamentalmente una entidad autónoma que percibe su entorno y actúa sobre él para alcanzar sus objetivos. Una definición clásica de Wooldridge y Jennings (1995) describe un agente inteligente como "un sistema informático situado en un entorno, capaz de actuar de forma flexible y autónoma para cumplir sus objetivos de diseño" (figura 1). Esto abarca varios aspectos importantes. El agente está situado en un mundo (físico o virtual) donde recibe entradas (percepciones) a través de sensores e influye en el mundo a través de actuadores. La autonomía implica que opera sin intervención humana directa, tomando sus propias decisiones sobre qué acciones realizar (Wooldridge, 2002). La acción flexible significa que el agente puede adaptarse a condiciones cambiantes y perseguir sus objetivos de manera robusta.

FIGURA 1

DEFINICIÓN DEL AGENTE DE WOOLDRIDGE Y JENNINGS



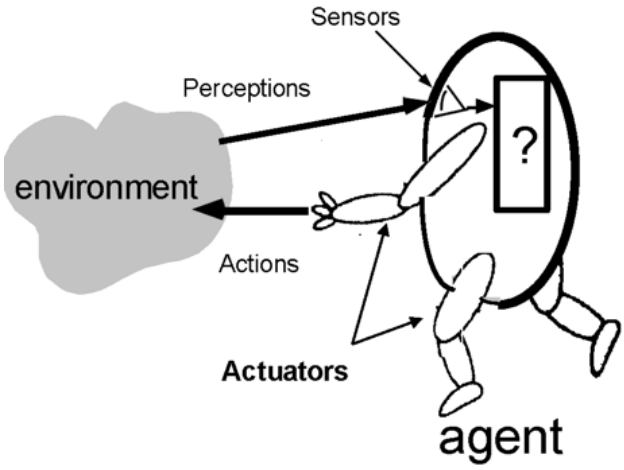
Fuente: Elaboración propia.

Wooldridge amplió que la acción flexible e inteligente implica propiedades como la reactividad, la proactividad y la capacidad social (Wooldridge & Jennings, 1995). Estas, junto con la autonomía, a menudo se citan como propiedades clave de los agentes inteligentes (Jennings *et al.,* 1998). La autonomía significa que los agentes funcionan de forma independiente, tomando decisiones y ejecutando acciones sin necesidad de supervisión humana directa; esta autonomía les permite realizar tareas y adaptarse en tiempo real a entornos cambiantes. La reactividad implica que el agente percibe su entorno y responde a los cambios de manera oportuna. La proactividad significa que puede tomar la iniciativa —un comportamiento orientado a objetivos— no solo reaccionando al entorno, sino también haciendo acciones dirigidas a objetivos para cumplir sus metas. La capacidad social se refiere a la habilidad de un agente para interactuar con otros agentes (o humanos), comunicándose, cooperando o compitiendo según sea necesario. Estos atributos distinguen a un agente inteligente de un mero programa que actúa de manera predefinida. Es importante destacar que estos conceptos se establecieron a mediados de la década de 1990, lo que subraya que el concepto de agentes de IA no es nuevo; los investigadores han estudiado durante mucho tiempo qué hace que un agente sea "inteligente".

Para ilustrar estas ideas, podemos visualizar un modelo simple de un agente inteligente interactuando con su entorno (figura 2). El agente recibe percepciones del entorno a través de sensores, procesa estas entradas (utilizando su razonamiento interno o mecanismo de toma de decisiones) y luego produce acciones a través de actuadores para cambiar el entorno. El ciclo percibir-pensar-actuar es el núcleo de cualquier agente. Por ejemplo, un agente termostato percibe la temperatura actual, "decide" si coincide con

FIGURA 2

MODELO SIMPLE DE UN AGENTE INTERACTUANDO CON SU ENTORNO



Fuente: Elaboración propia.

el punto de ajuste deseado y actúa encendiendo o apagando la calefacción. Un ejemplo más complejo es un agente robótico: podría percibir el mundo a través de cámaras y sensores de distancia, razonar sobre sus objetivos (por ejemplo, navegar a una ubicación) y actuar sobre su sistema de propulsión. En todos los casos, la noción de agencia implica que el sistema vincula continuamente la percepción con la acción en busca de objetivos.

El cuadro con el signo de interrogación en la figura 2 es donde incorporaremos la inteligencia del agente, es decir, aquellas técnicas de IA que nos permiten determinar, a partir de las percepciones del agente, qué acciones realizar en el entorno para alcanzar sus objetivos. Estas técnicas nos permitirán interpretar sus percepciones, deliberar para determinar qué objetivos cumplir, razonar para determinar qué acciones le permitirán alcanzar sus objetivos, etc. Esto implica el uso de técnicas de reconocimiento de formas, procesamiento del lenguaje natural, planificación, los LLM, etc.

El agente se concibe como una abstracción tecnológica independiente del mecanismo de razonamiento: puede basarse en lógica simbólica, planificación, sistemas basados en reglas, sistemas basados en casos, aprendizaje, redes neuronales, etc. Por tanto, los agentes actuales basados en LLM no rompen con este modelo, sino que lo amplían mediante nuevas formas de inferencia lingüística y contextual basadas en la tecnología de los LLM. La IA agentiva no introduce una nueva categoría ontológica, sino que reinterpreta la noción clásica de agente bajo un nuevo marco de implementación, centrado en la generación y comprensión del lenguaje natural. Los nuevos agentes basados en los LLM incorporan una técnica conocida como cadena

de pensamiento (CdP) razonamiento, para mejorar sus habilidades de resolución de problemas. Este método alienta a los modelos a dividir los problemas en pasos intermedios, imitando la forma en que los humanos piensan en un problema de manera lógica esto es algo que ya se propuso en la inteligencia artificial distribuida. Esto no supone un razonamiento genuino, como lo entendemos en IA; se asemeja más a un proceso de resolución de problemas de búsqueda en un espacio de soluciones, una técnica clásica de la IA.

Vale la pena señalar que el libro de texto de IA de Russell y Norvig (2020) identifica esta interacción agente-entorno como el concepto fundamental de la IA, incluso definiendo la IA misma como el estudio de los agentes inteligentes. A finales de la década de 1990, el paradigma de agente se había vuelto central en la investigación de la IA, dando lugar a subcampos como los agentes autónomos y los sistemas multiagente. Esto llevó a la creación de espacios dedicados como el International Workshop on Agent Theories, Architectures, and Languages (ATAL), la International Conference on Autonomous Agents (AGENTS), la International Conference on Multi-Agent Systems (ICMAS) y su fusión en la Autonomous Agents and Multiagent Systems Conference (AAMAS) a partir de 2002. Sociedades profesionales como The International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS) y The European Association for Multi-Agent Systems (EURAMAS), The Foundation for Intelligent Physical Agents (FIPA) y revistas como Autonomous Agents and Multi-Agent Systems (Springer) se establecieron para avanzar en la teoría y práctica de los agentes. El área de agentes y sistemas multiagente ha estado presente durante muchos años en conferencias de inteligencia artificial, en particular en las más prestigiosas como IJCAI (International Joint Conference on Artificial Intelligence), AAAI (Association for the Advancement of Artificial Intelligence) y ECAI (European Conference on AI). En resumen, para cuando los blogs tecnológicos actuales comenzaron a hablar de "Agentic AI", la comunidad de investigación de IA ya había construido un campo vibrante y en evolución en torno a estos conceptos.

La literatura distingue entre tres niveles analíticos (Ferber, 1999): conceptual, arquitectónico y tecnológico. El nivel conceptual define las propiedades esenciales del agente; el arquitectónico describe su estructura interna (reactiva, deliberativa o híbrida); y el tecnológico implementa dichas funciones mediante diversas técnicas. La IA agentiva opera principalmente en este último nivel, reemplazando/complementando el razonamiento simbólico o el razonamiento práctico (Bratman, 1999 [1987]) por modelos de lenguaje que permiten deliberación en lenguaje natural y adaptabilidad contextual. El nivel conceptual sigue siendo el mismo, aunque el soporte computacional evolucione.

4. SISTEMAS MULTIAGENTE

Aunque un solo agente inteligente puede ser poderoso, muchos problemas del mundo real requieren múltiples agentes que cooperen (o compitan) para ser resueltos. Un sis-

tema multiagente consta de múltiples agentes autónomos que interactúan —ya sea de forma colaborativa o competitiva— para alcanzar objetivos individuales o colectivos (Wooldridge, 2002). En tales sistemas, ningún agente individual tiene capacidades suficientes para lograr los objetivos generales por sí solo; esta interacción puede implicar comunicación, negociación, cooperación en subtareas o competencia por recursos compartidos, dependiendo del escenario.

A principios de la década de 2000, los SMA habían madurado como un área de investigación distinta, abordando los desafíos de coordinación, comunicación y organización de múltiples agentes. Los investigadores desarrollaron estándares y lenguajes para permitir la interoperabilidad entre agentes de diferentes desarrolladores. Por ejemplo, la Fundación para Agentes Físicos Inteligentes (FIPA) comenzó a emitir estándares para protocolos de comunicación de agentes en 1996 (FIPA, 1996).

Esta área de investigación se basa en un nuevo paradigma de computación, la computación como interacción (Shoham, 1993). En este paradigma, la computación ocurre a través de la comunicación entre entidades computacionales; la computación es una actividad inherentemente social, no solitaria, lo que da lugar a nuevas formas de concebir, diseñar, desarrollar y gestionar sistemas computacionales.

Cuando hablamos de comunicación, no podemos olvidar iniciativas como el Esfuerzo de Compartición de Conocimiento (*Knowledge Sharing Effort*). Esta iniciativa fue lanzada alrededor de 1990 por ARPA (la Asociación de Proyectos de Investigación Avanzada del Departamento de Defensa de EE. UU.), con el apoyo de organizaciones de investigación como ASOFR, NSF y NRI. Su objetivo principal era promover la interoperabilidad en la comunicación de conocimiento entre sistemas inteligentes.

Como resultado de este proyecto, se concluyó que la comunicación efectiva entre agentes distribuidos requiere tres niveles de lenguaje:

1. *Sintaxis*: Se requiere un lenguaje para representar el conocimiento de manera formal y estructurada.

2. *Semántica*: Se requiere un lenguaje para definir ontologías, lo que permite describir el significado del conocimiento representado.

3. *Pragmática*: Se requiere un lenguaje para la comunicación entre agentes, centrándose en cómo se intercambian y manipulan los mensajes de conocimiento.

El diseño de estos lenguajes de comunicación de agentes se inspiró en teorías lingüísticas sobre los actos de habla (Austin, 1962; Searle, 1969):

- J. L. Austin (1962) – Cómo hacer cosas con palabras
- J. R. Searle (1969) – Actos de habla

Estos trabajos sentaron las bases para comprender la pragmática de la comunicación, es decir, cómo se utilizan las palabras para realizar acciones.

Los primeros lenguajes de comunicación de agentes se basaron en la teoría de los actos de habla de la filosofía del lenguaje. Concretamente, se propusieron lenguajes como KQML (*Knowledge Query and Manipulation Language*) (Finin *et al.,* 1994) y FIPA-ACL (FIPA, 1996) para estandarizar cómo los agentes hacen preguntas, se informan mutuamente o negocian. También se desarrollaron formatos comunes de representación del conocimiento, como KIF (*Knowledge Interchange Format*) (Genesereth & Fikes, 1992) para la sintaxis del contenido y lenguajes como ONTOLINGUA (Ontolingua, s.f.) para definir ontologías compartidas para la semántica. Estos esfuerzos fueron el intento de la comunidad de SMA de garantizar que agentes heterogéneos pudieran trabajar juntos, un reconocimiento de que la capacidad social (una de las propiedades de los agentes) requería métodos de comunicación estandarizados.

En los sistemas multiagente, los agentes que los constituyen deben interactuar para resolver el problema para el que fueron diseñados; estas interacciones pueden ser colaborativas, cuando los agentes son benevolentes, o competitivas, cuando los agentes son egoístas. Esto se traduce en la necesidad de métodos que permitan a los agentes llegar a acuerdos o alinear sus intenciones, lo que lleva a conceptos como consenso distribuido, trabajo en equipo o subastas para la asignación de tareas (*Agreement Computing)* (Sierra *et al.*, 2011).

En respuesta a esta necesidad de alcanzar acuerdos, surgieron las tecnologías de acuerdo (*Agreement Technologies*, AT) (Ossowski *et al.*, 2012). Estas se refieren a sistemas informáticos en los que agentes de *software* autónomos negocian entre sí, generalmente en nombre de personas, para llegar a acuerdos mutuamente aceptables. La autonomía, la interacción, la movilidad y la apertura son características relevantes desde una perspectiva teórica y práctica.

En Europa, por ejemplo, a finales de la década de 2000 se lanzó un gran programa de investigación sobre tecnologías de acuerdo (*Agrement Techologies-Cost Action*) para investigar estos temas. Todo esto subraya que, para 2010, los problemas de coordinación de múltiples actores de IA —desde la comunicación y la cooperación hasta la gobernanza— ya estaban siendo estudiados activay parcialmente resueltos.

En resumen, la comunidad de IA ha tenido durante mucho tiempo definiciones claras de lo que es un agente de IA y lo que es un sistema multiagente. Estas definiciones

giran en torno a la autonomía, la acción intencional y, para los SMA, la interacción entre agentes. A continuación, resumimos las características clave de cada concepto:

- Agente inteligente: una entidad autónoma y flexible situada en un entorno, con propiedades de autonomía, reactividad, proactividad y capacidad social.

- Sistema multiagente: una colección de agentes autónomos que interactúan en un entorno común, posiblemente cooperando o compitiendo, lo que requiere mecanismos de comunicación y coordinación.

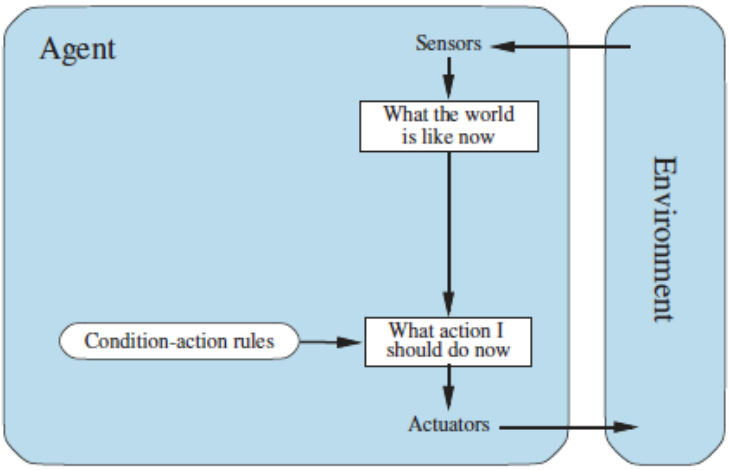
A partir de estos conceptos, se puede deducir que esta es un área bien establecida en el campo de la IA. A continuación, nos centraremos en cómo se construyen los agentes de IA, es decir, las arquitecturas que permiten a los agentes exhibir un comportamiento inteligente.

5. ARQUITECTURAS DE AGENTES INTELIGENTES

Stuart Russell y Peter Norvig, en su libro *Inteligencia Artificial: Un Enfoque Moderno* (Russell & Norvig, 2020) clasifican los agentes inteligentes según su capacidad para percibir, modelar, planificar y aprender. La evolución va desde simples agentes reactivos hasta agentes con capacidades de aprendizaje autónomo, y definen cinco arquitecturas de agentes abstractas que describimos brevemente a continuación.

FIGURA 3

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

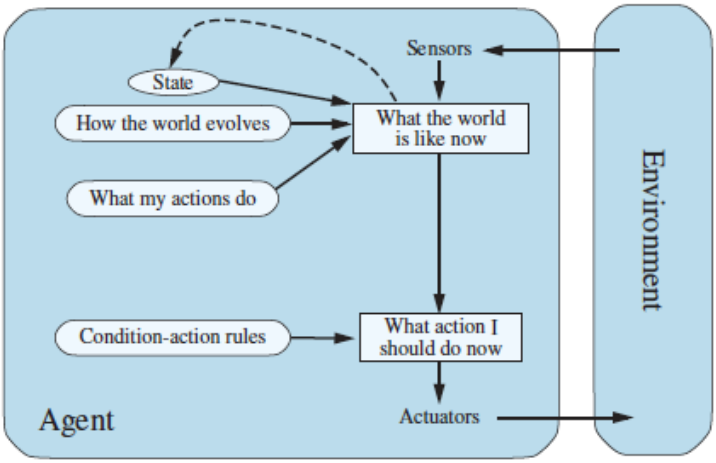
Agentes reactivos simples: Actúan directamente basándose en las percepciones actuales, utilizando reglas de condición → acción. No tienen memoria ni un modelo interno del entorno.

Agentes basados en modelos (memoria): Mantienen un modelo interno para representar el estado del entorno. Actualizan este modelo basándose en las percepciones recibidas.

Agentes basados en objetivos: Seleccionan acciones basadas en objetivos específicos. Usan técnicas de búsqueda y planificación para alcanzar estos objetivos.

FIGURA 4

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

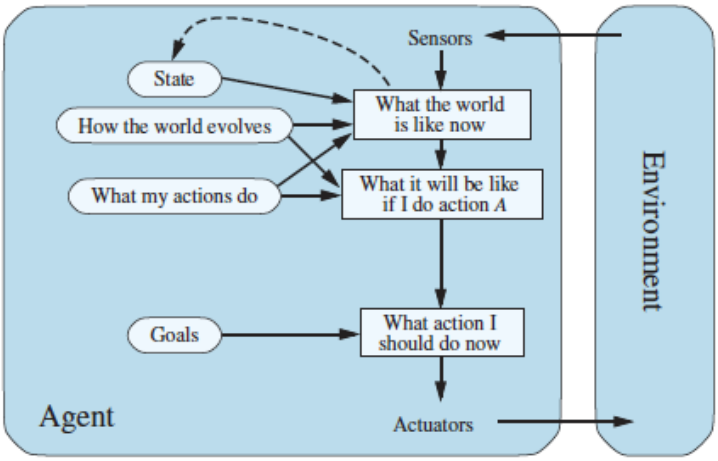
Agentes basados en utilidad: Generalizan los agentes basados en objetivos utilizando una función de utilidad para evaluar diferentes resultados y seleccionar el más preferido.

Agentes de aprendizaje: Capaces de mejorar su rendimiento a lo largo del tiempo a través de la experiencia. Incluyen módulos de acción, crítica y aprendizaje.

Diseñar un agente inteligente implica especificar cómo procesa la información desde la percepción hasta la acción. A lo largo de los años, los investigadores de IA han propuesto diversas arquitecturas de agentes, es decir, modelos subyacentes que definen la estructura interna y los procesos de toma de decisiones de un agente. Van desde los sistemas reactivos simples a los deliberativos complejos. A continuación, repasamos

FIGURA 5

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



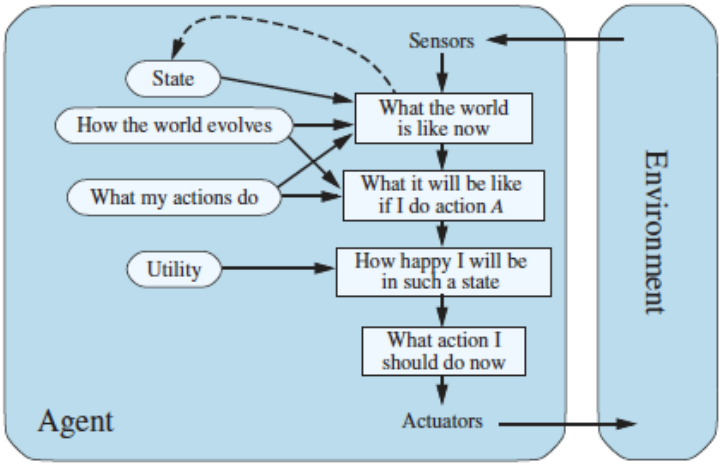
Fuente: Russell & Norvig (2020).

varios de los principales paradigmas arquitectónicos, ya que su comprensión aclarará cómo encajan (o no) los llamados sistemas "agentivos" con los enfoques tradicionales.

Agentes reactivos (reflejos): El tipo más simple de arquitectura de agente es un agente reactivo, también llamado agente reflejo simple. Este agente selecciona acciones basa-

FIGURA 6

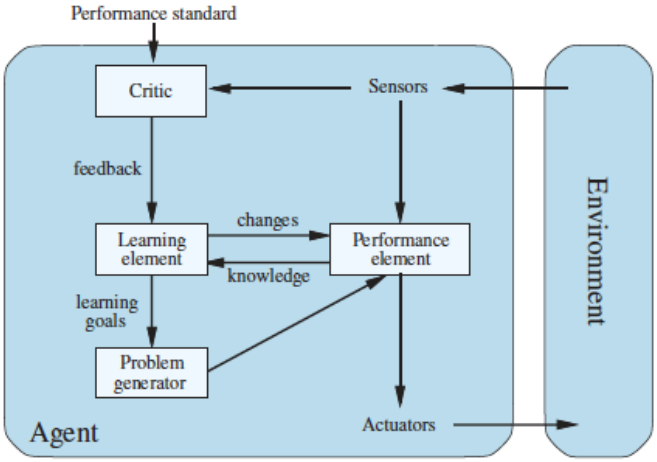
ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

FIGURA 7

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

das en la percepción actual, ignorando la historia perceptiva. Implementa un mapeo directo de situaciones a acciones, básicamente un conjunto de reglas de condición-acción. Si la condición coincide con el estado actual del entorno, el agente ejecuta la acción asociada; si ninguna regla coincide, no hace nada. Los agentes reactivos no planifican explícitamente ni modelan el mundo, sino que actúan en tiempo real siguiendo una lógica "si-entonces". Este paradigma cobró importancia a mediados de los 80 como reacción a las limitaciones de la planificación simbólica pura en IA. La arquitectura de subsunción de Rodney Brooks (Brooks, 1986) es un ejemplo. En este modelo, el sistema de control del agente se construye en capas de comportamiento (por ejemplo, evitación de obstáculos, deambulación, búsqueda de objetivos), donde los comportamientos reactivos de bajo nivel pueden anular a los de alto nivel cuando sea necesario. Esto tuvo una gran influencia en la robótica basada en el comportamiento, ya que demostró que pueden surgir comportamientos eficaces a partir de redes de módulos simples y reflejos, sin un razonamiento simbólico centralizado. La ventaja de los agentes reactivos es su velocidad y robustez en entornos dinámicos (sin deliberaciones complejas que los ralenticen); la desventaja es que los agentes puramente reactivos tienen dificultades con los comportamientos estratégicos a largo plazo, ya que carecen de un modelo interno del mundo o de objetivos explícitos.

Agentes deliberativos (simbólicos): En el otro extremo del espectro están los agentes deliberativos, a veces llamados agentes de razonamiento deductivo (especialmente en la literatura temprana). Estos agentes mantienen un modelo simbólico del mundo y utilizan el razonamiento simbólico (lógica) para decidir sus acciones. En una arquitectura deliberativa, el agente tiene representaciones explícitas de creencias sobre el mundo y

utiliza la deducción lógica o la búsqueda para idear un plan de acción que alcance sus objetivos. Este enfoque se remonta a los inicios de la IA (años 50) y continuó en los años 70 y 80 en sistemas que intentaban demostrar teoremas o deducir planes a partir de principios formales. Un agente puramente deductivo podría, por ejemplo, utilizar la demostración automática de teoremas para deducir la mejor acción dado un objetivo formal y una base de conocimiento formal, un método que es correcto por diseño pero a menudo inviable desde el punto de vista computacional en entornos complejos y dinámicos. A finales de la década de 1980, se reconoció que los agentes puramente simbólicos eran frágiles y lentos en la práctica; los entornos de la vida real cambian más rápido de lo que pueden replanificarse, y el problema del encuadre (determinar qué cambia y qué permanece igual después de una acción) hace que los enfoques lógicos simples sean poco prácticos. Esta constatación dio lugar al movimiento de agentes reactivos antes mencionado, así como a arquitecturas que hibridan ambos extremos. Ejemplos de este tipo de arquitectura son AGENT0 y PLACA.

Agentes híbridos: Las arquitecturas híbridas intentan combinar los puntos fuertes de los enfoques reactivo y deliberativo. Un patrón común es un diseño de dos (o tres) capas: una capa reactiva se encarga de las respuestas de bajo nivel y en tiempo crítico al entorno, mientras que una capa deliberativa se ocupa de la planificación de alto nivel y la gestión de objetivos. Las dos capas deben coordinarse; por ejemplo, la capa reactiva puede anular los planes para hacer frente a amenazas inmediatas o, a la inversa, la capa deliberativa puede suprimir algunos comportamientos reactivos cuando se persigue un plan a largo plazo. Una de las primeras arquitecturas híbridas famosas fue la arquitectura TouringMachines (Ferguson, 1992), que tenía capas para las habilidades reactivas, la planificación y una capa de supervisión para gestionar los conflictos. Otros ejemplos son InteRRaP y las arquitecturas 3T (de tres niveles). En la década de 1990, el diseño híbrido de agentes era la norma para los robots complejos y los agentes de *software*: reconocía que ningún método es suficiente para todos los aspectos de la inteligencia. El enfoque híbrido reconoce la necesidad de reactividad para manejar los cambios en tiempo real y de deliberación para manejar el razonamiento estratégico. Muchos agentes de IA modernos siguen implícitamente este enfoque por capas, aunque no estén explícitamente etiquetados como tales.

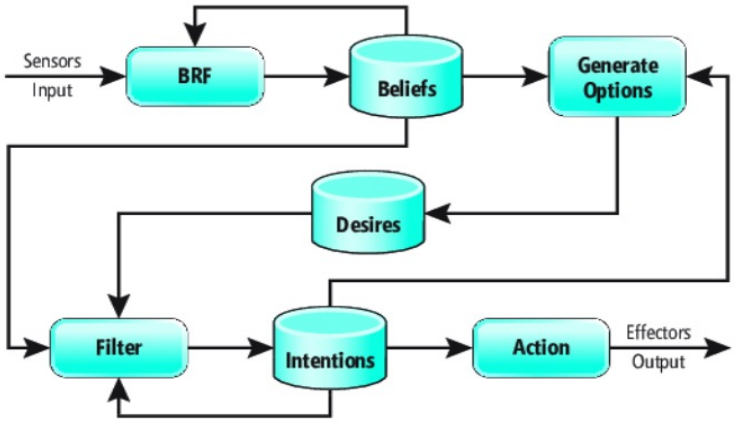
Agentes de Creencia-Deseo-Intención (Belief-Desire- Intention BDI): Una arquitectura particularmente importante, que se origina en el trabajo de Michael Bratman (Bratman, 1999 [1987]) sobre el razonamiento práctico humano y fue adaptada para la IA en la década de 1990. La arquitectura BDI proporciona un marco teórico estructurado para la deliberación del agente, está definida por los siguientes conceptos:

- Creencias (B): el estado de información del agente, lo que cree sobre el mundo.
- Deseos (D): los objetivos o situaciones que al agente le gustaría lograr.

- Intenciones (I): el subconjunto de deseos con los que el agente se ha comprometido.

FIGURA 8

ESQUEMA DE LA ARQUITECTURA DE UN AGENTE BDI



Nota: Esquema de la arquitectura de un agente BDI, que muestra cómo las percepciones del entorno son procesadas por la función de revisión de creencias (BRF) para actualizar las creencias del agente, que junto con las intenciones existentes informan la generación de nuevas opciones (posibles deseos). A continuación, un filtro selecciona qué deseos adoptar como Intenciones actuales, y el agente actúa para lograr estas intenciones. Este ciclo de observación, actualización de creencias, deliberación (opciones → intenciones) y acción se repite continuamente.

Fuente: https://learn.microsoft.com/en-us/archive/msdn-magazine/2019/january/machine-learning-leveraging-the-beliefs-desires-intentions-agent-architecture

El modelo BDI define un bucle de control en el que el agente actualiza continuamente sus creencias, genera opciones (deseos) basadas en esas creencias y en las intenciones actuales, filtra (delibera) esos deseos para elegir con cuáles comprometerse (intenciones) y, a continuación, planifica para determinar qué acciones desempeñar para satisfacer las intenciones y, finalmente, actúa para cumplir las intenciones. Esto se asemeja mucho a la forma en que los humanos razonamos sobre las acciones: tenemos muchos deseos potenciales en cualquier momento, pero los filtramos en función de las prioridades y la viabilidad para convertirlos en intenciones, que luego intentamos ejecutar. Un agente BDI no vuelve a planificar a ciegas desde cero en cada momento, sino que mantiene su compromiso con las intenciones hasta que se cumplen determinadas condiciones (objetivo alcanzado, objetivo imposible o sustituido por un objetivo de mayor prioridad), lo que proporciona un equilibrio entre la persistencia proactiva y la adaptabilidad reactiva.

La arquitectura BDI formaliza el razonamiento práctico en los agentes, permitiéndoles equilibrar la reactividad (mediante actualizaciones perceptivas) con el comportamiento orientado a objetivos (mediante intenciones). Se ha utilizado como base para muchos lenguajes y plataformas de programación de agentes (por ejemplo, PRS, JACK, JAM). El énfasis del modelo BDI en el estado mental del agente (actitudes informativas y motivacionales) conecta con la teoría de sistemas intencionales de Dennett, proporcionando una forma rigurosa de diseñar agentes que parecen tener creencias, deseos e intenciones; de hecho, en BDI, estas son estructuras de datos explícitas.

Además de las arquitecturas presentadas, existen otras muchas (agentes impulsados por eventos, agentes de redes neuronales, etc.), pero las comentadas ilustran los principales paradigmas. Está claro que, a principios de la década de 2000, los investigadores de IA habían desarrollado un sofisticado conjunto de herramientas para construir agentes. Las arquitecturas reactivas ofrecían velocidad, las deliberativas previsión, las híbridas intentaban combinar lo mejor de ambas y las BDI añadían claridad filosófica y estructura práctica a las estrategias de compromiso.

Es en este contexto en el que consideramos la reciente oleada de "Agentic AI". Cuando los comentarios actuales hablan de sistemas de IA agentiva, debemos preguntarnos: ¿cuál de estas arquitecturas establecidas (o combinaciones) se está utilizando realmente? ¿Los actuales "agentes autónomos basados en los LLM" representan algo fundamentalmente novedoso, o podrían ser una implementación de, por ejemplo, un ciclo BDI con un LLM que se encarga de parte de la toma de decisiones? Para explorar esta cuestión, examinamos ahora qué significa la Agentic AI en el contexto de las tendencias actuales de la IA generativa.

6. DE LOS LLM A LA "AGENTIC AI": LA NUEVA OLA DE AGENTES AUTÓNOMOS

En el año 2023 se produjo una explosión de interés por encadenar y orquestar grandes modelos lingüísticos para realizar tareas autónomas orientadas a objetivos. El lanzamiento de potentes LLM (GPT-3, GPT-4, GPTo, Gemini, Deepsek, Llama, etc.) capaces de razonar y generar código llevó a los desarrolladores a crear sistemas como AutoGPT, BabyAGI, LangChain, GraphChain, AutoGen, CAMEL, ADK y muchos otros, que han permitido la creación de agentes y SMA en los que los agentes son LLM capaces de razonar, planificar y comunicarse entre sí utilizando lenguaje natural. Estos SMA basados en LLM aprovechan la inteligencia colectiva de múltiples agentes especializados para abordar tareas complejas de forma más eficiente que un único agente. Un conjunto de LLM puede asumir funciones especializadas definidas por instrucciones, colaborando mediante el intercambio de información y la coordinación de acciones en lenguaje natural, de forma similar a la interacción humana. Sin embargo, el éxito de los LLM en tareas complejas requiere una intervención humana considerable para guiar la genera-

ción de textos y la toma de decisiones. Por tanto, la coordinación autónoma entre LLM plantea varios retos, como garantizar que los agentes cooperen de forma que se alineen con los objetivos humanos. Otro hito importante es que se han desarrollado también propuestas de estandarización como MCP (*Model Context Protocol*) desarrollado por Anthropic (estándar abierto que permite a los desarrolladores crear conexiones seguras y bidireccionales entre sus fuentes de datos y herramientas basadas en inteligencia artificial) o A2A (Agent2Agent Protocol, un estándar abierto que permite a los agentes de IA comunicarse y colaborar entre diferentes plataformas y marcos, independientemente de las tecnologías subyacentes). Aunque en esta nueva ola de los agentes basados en LLM aún queda mucho trabajo de estandarización, que debería basarse en los trabajos ya realizados en el área de los sistemas multiagente, son iniciativas importantes que van en el camino adecuado.

Esta tendencia ha ido acompañada de la popularización del término "Agentic AI (IA agentiva)". Pero, ¿qué significa exactamente este término?

En el ámbito industrial, el término "Agentic AI" (Softude, 2025; Silfverskiöld, 2025) se refiere generalmente a los sistemas de IA que pueden decidir y ejecutar acciones de forma autónoma para alcanzar un objetivo de alto nivel, a menudo dividiendo el objetivo en subtareas y posiblemente coordinando múltiples componentes o subagentes. Implica un grado de iniciativa (proactividad) y planificación a largo plazo que va más allá de las respuestas puntuales o la generación de texto. Por ejemplo, un asistente de Agentic AI no solo podría responder a una consulta del usuario, sino también tomar medidas proactivas (por ejemplo, programar una reunión, enviar correos electrónicos, ejecutar transacciones) sin indicaciones explícitas para cada paso. En esencia, esto describe un agente inteligente en el sentido clásico. El agente percibe un objetivo, decide de forma iterativa "¿Qué subobjetivo o acción debe realizar a continuación para alcanzarlo?", posiblemente invoca herramientas externas o API, interactúa con otros agentes, supervisa el progreso, etc. El entusiasmo en torno a estas capacidades en 2023-2024 es real, impulsado por impresionantes demostraciones de agentes basados en LLM, pero conceptualmente, esto puede identificarse claramente con arquitecturas y conceptos de agentes autónomos que existen desde hace décadas.

Es revelador ver cómo las fuentes modernas distinguen entre "Agentic AI" y "agentes de IA". Según un artículo de IBM (IBM, 2023), "la Agentic AI es el framework; los agentes de IA son los componentes básicos dentro de ese framework". En otras palabras, la Agentic AI se refiere a un sistema global (quizás un ecosistema de módulos que incluye uno o más agentes) diseñado para resolver problemas con una supervisión mínima, mientras que un agente de IA es un componente autónomo específico que realiza tareas individuales dentro de dicho sistema. Softude (2025) describe de manera similar a un agente de IA como "un bloque de construcción de Agentic AI... programado para funcionar de forma autónoma", con bucles de percepción, decisión y acción, mientras que la Agentic AI se refiere a sistemas con un alto grado de agencia que "establecen

objetivos de forma proactiva, toman decisiones y actúan con una intervención humana mínima". Los sistemas denominados de IA agentiva implementan los principios clásicos de agencia usando modelos fundacionales. Los agentes basados en los LLM ejecutan tareas complejas mediante razonamiento lingüístico, planificación y acceso a herramientas externas. Esta tendencia no constituye un nuevo paradigma, sino una ampliación tecnológica del marco deliberativo tradicional. En este sentido, la IA agentiva se podría entender como una instancia moderna del modelo BDI (*Belief–Desire–Intention*) (Jao & Georgeff, 1991 y 1995), donde las creencias son *embeddings* contextuales, los deseos son prompts y las intenciones, secuencias de acciones generadas.

La clave es que la Agentic AI a menudo implica no un solo agente, sino un conjunto orquestado de agentes y herramientas, junto con una arquitectura que les permite coordinarse dinámicamente. En términos de computación en la nube, un sistema de Agentic AI podría considerarse como un conjunto o flujo de trabajo (a veces llamado un "sistema de orquestación de agentes"), mientras que un agente inteligente es un actor autónomo individual dentro de él.

En IBM (2023), el ejemplo de IBM para la IA agencial es un sistema inteligente de energía doméstico que gestiona múltiples agentes dispositivos (IoT). La Agentic AI (el sistema global) comprende el objetivo general del propietario (eficiencia energética) y coordina los agentes de IA individuales —el agente termostato, el agente iluminación, los agentes electrodomésticos—, cada uno con su propia tarea y objetivo específicos. Este ejemplo se corresponde con el concepto de sistema multiagente: el hogar inteligente tiene múltiples agentes (termostato, luces, etc.) que, en este caso, cooperan entre sí. De hecho, no se introduce nada conceptualmente exótico al llamarlo "agente": se trata esencialmente de un sistema de control multiagente inteligente, un tema ampliamente tratado en la investigación sobre IA distribuida.

Entonces, ¿por qué esta nueva terminología? Probablemente se deba en parte al *marketing* o al ciclo natural de las ideas que son redescubiertas por nuevas comunidades. El resurgimiento del interés por los "agentes inteligentes" entre los desarrolladores se produjo tras la llegada de los LLM avanzados, que proporcionaron un nuevo conjunto de herramientas para implementar comportamientos de agentes (por ejemplo, utilizando LLM para interpretar objetivos y generar planes en lenguaje natural). Para comercializar esto, términos como "Agentic AI" ganaron popularidad, tal vez para señalar un cambio de los modelos de IA estáticos (que generan resultados de forma pasiva) a los sistemas de IA que son participantes activos y orientados a objetivos (agentes). Sin embargo, lo irónico es que las nociones subyacentes de autonomía e interacción multiagente no son nada nuevas.

Como se ha comentado anteriormente, también hemos asistido a una proliferación de plataformas y bibliotecas de código abierto para facilitar la construcción de estos agentes basados en LLM. De hecho, en 2024-2025, han surgido docenas de nuevas

plataformas basadas en LLM, muchos de ellas lanzadas en el plazo de un año. Algunos ejemplos son AutoGen, AgentScope, CrewAI, CamelAI, ADK, OpenAI, 8n8, LangChain, AutoGen, Hugging Face Transformers Agents, LangGraph, Agenta, Minichain, Auto-GPT, BabyAGI y muchos más. Estas plataformas integran modelos LLM que permiten el desarrollo de agentes basados en LLM para razonar, planificar y actuar en entornos abiertos. Su funcionamiento se basa en tres capacidades clave:

- Razonamiento: Los LLM generan cadenas de pensamiento para razonar antes de actuar.
- Memoria: Integran recuerdos episódicos y a largo plazo (almacenes vectoriales, registros) para contextualizar las decisiones.
- Uso de herramientas: Utilizan herramientas externas (calculadoras, navegadores, API) a través de indicaciones estructuradas.

Proporcionan abstracciones para definir herramientas, gestionar instrucciones y conectar múltiples agentes basados en LLM.

Plataformas como LangChain, AutoGen y CrewAI obtuvieron rápidamente decenas de miles de estrellas en GitHub (barras doradas), lo que indica un gran interés por parte de los desarrolladores, mientras que muchas plataformas más recientes entraron en escena a finales de 2024 con una creciente aceptación. Esto refleja una rápida expansión de las herramientas y bibliotecas para crear agentes inteligentes autónomos, a menudo centrados en la orquestación de LLM con herramientas externas. La tendencia subraya el entusiasmo de la industria por las soluciones de IA basadas en agentes, pero también conlleva el riesgo de fragmentación y repetición de ideas conocidas en la investigación académica sobre agentes.

La proliferación de marcos sugiere que muchos se enfrentan a retos similares: cómo permitir que un agente de IA utilice herramientas (API, bases de datos), cómo mantener la memoria de interacciones pasadas, cómo coordinar múltiples agentes o encadenar subtareas, etc. Se trata, en esencia, de implementaciones de lo que la comunidad de investigación sobre agentes reconocería como planificación, memoria/base de conocimientos y coordinación multiagente. Por ejemplo, un patrón de diseño común es un agente "planificador" basado en LLM que puede generar agentes "ejecutores" para subtareas, que luego informan de sus resultados, lo que recuerda a la planificación jerárquica de tareas o a las organizaciones de agentes maestro-esclavo de la literatura sobre MAS.

Otro ejemplo: el concepto de dar acceso a un agente LLM a herramientas (como búsquedas web o calculadoras) y esperar que decida cuándo usar cada una de ellas, es análogo a un agente que razona sobre sus acciones y elige un efector o subrutina apro-

piados. Clásicamente, esto podría asignarse al repertorio de acciones del agente y al acceso al entorno. Hoy en día, los investigadores están inventando técnicas de instrucción (por ejemplo, "ReAct", que entrelaza el razonamiento con el uso de herramientas) para lograr estos comportamientos con los LLM, de forma similar a trabajos anteriores sobre sistemas de planificación de agentes que incluían acciones tanto internas (razonamiento) como externas (ejecución de una llamada API).

El verdadero valor añadido de los nuevos sistemas basados en LLM reside en la flexibilidad y los conocimientos que encapsulan los grandes modelos preentrenados. El hecho de que una instancia de GPT-4, Gemini, Llama, Deepsek o cualquier otro modelo pueda analizar una instrucción arbitraria, descomponerla en pasos y generar código o texto para llevarlos a cabo, supone un gran avance en cuanto a capacidad en comparación con, por ejemplo, un agente BDI codificado de forma rígida que solo conoce una biblioteca fija de planes. Esto ha permitido crear agentes mucho más polivalentes. Por ejemplo, un LLM bien diseñado puede servir como interfaz universal entre los objetivos del lenguaje natural y las acciones formales, funcionando como el "cerebro" del agente que razona a un alto nivel de forma abierta. Esto no era posible con una buena fiabilidad antes del aprendizaje profundo, por lo que el entusiasmo en torno a la Agentic AI está justificado: estos sistemas pueden hacer cosas con las que los sistemas de agentes anteriores tenían dificultades, gracias a la potencia de los LLM.

La velocidad con la que se están creando herramientas de desarrollo de sistemas de agentes y multiagente basados en LLM es también un factor tremendamente positivo. Durante la evolución del campo de los agentes y los sistemas multiagentes, la disponibilidad de plataformas de desarrollo (JADE, JACK, MADKit, ZEUS, Magentix, etc.) evolucionó muy lentamente, mientras que hoy en día tenemos una amplia gama entre la que elegir, y siguen aumentando.

Sin embargo, a nivel conceptual, los sistemas de Agentic AI siguen siendo agentes inteligentes. Del mismo modo, un sistema multiagentic (término que se utiliza ocasionalmente para referirse a escenarios multiagente LLM) es, por definición, un sistema multiagente.

7. MAL USO DE LA TERMINOLOGÍA Y CONFUSIÓN CONCEPTUAL

De lo que hemos discutido hasta ahora, podemos abordar el meollo del asunto: el mal uso de los términos "agentic (agentiva)" y "multiagentic (multiagentiva)" en el discurso actual sobre la IA. Muchos artículos y comunicados de prensa recientes utilizan estas palabras como si denotaran una nueva tecnología, cuando en realidad están describiendo capacidades o diseños de sistemas examinados durante mucho tiempo bajo las banderas de los agentes inteligentes y los sistemas multiagente. Esta confusión conceptual puede ser problemática por varias razones:

- Oculta la riqueza de la investigación existente. Al no reconocer que la Agentic AI se refiere esencialmente a agentes autónomos, los nuevos profesionales pueden pasar por alto décadas de literatura sobre cómo diseñar y evaluar estos sistemas. Como resultado, pueden repetir errores del pasado o volver a trabajar en problemas ya resueltos. De hecho, una de las principales motivaciones de este artículo es la preocupación de que ignorar el trabajo anterior nos lleve a "reinventar la rueda" al aplicar la Agentic AI a nuevos ámbitos.

- Crea confusión terminológica. Tradicionalmente, en IA, el adjetivo "agentic" rara vez se utilizaba; el término "agencia" ha sido más frecuente; los investigadores simplemente hablaban de sistemas basados en agentes, agentes autónomos, etc. Ahora vemos frases como "agentic AI agents", que son redundantes (literalmente, "agentes de IA similares a agentes"). El término "multiagentic" es aún más problemático, ya que intenta adjetivar "multiagente". Decir "sistema multiagentic" es un neologismo innecesario cuando el término "sistema multiagente" es claro y está bien definido. Esta jerga puede confundir a los recién llegados y enturbiar las comunicaciones, especialmente entre la industria y el mundo académico. Por ejemplo, un ingeniero de *software* podría decir: "Hemos creado una IA multiagente para hacer X", mientras que un investigador de IA lo interpretaría como "un sistema multiagente para X". Es de esperar que se refieran a lo mismo, pero la terminología divergente puede impedir la transferencia de conocimientos.

- Puede exagerar la novedad. Cuando se acuña un nuevo término, a veces se da a entender que el concepto en sí es nuevo. Esto puede dar lugar a expectativas exageradas o a un reconocimiento injustificado. Consideremos que afirmaciones como "La Agentic AI va un paso más allá al permitir la toma de decisiones y la ejecución de tareas de forma autónoma", es una descripción acertada, pero que se aplica igualmente a los agentes inteligentes desde la década de 1990. O afirmaciones como "La Agentic AI se refiere a sistemas en los que colaboran múltiples agentes de IA... A diferencia de un único agente inteligente, que opera de forma aislada", básicamente describe un sistema multiagente. Sin contexto histórico, sin rigor científico, los lectores pueden pensar que estas capacidades fueron inventadas por los laboratorios de IA actuales.

Seamos claros: el término "agentic" no es un concepto nuevo en la IA, tiene raíces en las ciencias sociales y cognitivas y fue adoptado posteriormente (moderadamente) en la IA, pero las ideas subyacentes han sido parte del desfile de agentes de la IA durante mucho tiempo. Cuando los documentos técnicos de 2023 hablan de "Agentic AI", generalmente se refieren a lo que los investigadores de IA llamarían simplemente un agente inteligente/agente autónomo o un sistema basado en agentes. En otras palabras, deberíamos llamar a las cosas por su nombre: si es un agente de IA, llámalo

agente; si es un sistema multiagente, llámalo así. La IA agentiva no sustituye a los agentes autónomos/ sistemas multiagente, sino que los reinterpreta como ecosistemas de agentes lingüísticos, donde la interacción se da en lenguaje natural y la coordinación se logra mediante mecanismos de argumentación y diálogo.

Para ser claros, no hay que menospreciar o restar importancia a los recientes logros de la ingeniería en la implementación de agentes basados en LLM con nuevas herramientas de IA. Los avances son reales, pero encajan dentro de un marco conceptual ya existente. Al reconocer esto, los desarrolladores e investigadores pueden beneficiarse enormemente de las lecciones ya aprendidas. Por ejemplo:

La comunidad MAS ha estudiado los comportamientos emergentes en los sistemas multiagente. Algunos de los primeros experimentos con sistemas multiagente basados en LLM han observado comportamientos sorprendentes (positivos o negativos) cuando varios agentes LLM interactúan (por ejemplo, quedarse atascados en bucles de preguntas entre ellos). Muchos de ellos podrían entenderse utilizando enfoques de teoría de juegos o de coordinación desarrollados para MAS.

Las cuestiones relacionadas con la confianza y la seguridad en los agentes autónomos (como que un agente de IA se vuelva incontrolable o cause daños no deseados) se han debatido durante mucho tiempo en el contexto de la autonomía limitada y la autonomía ajustable. Existen marcos para incorporar restricciones o normas éticas en la toma de decisiones de los agentes. En lugar de empezar desde cero, las nuevas iniciativas de Agentic AI pueden adaptar estos marcos a los agentes basados en LLM.

La importancia de los estándares no puede subestimarse. A finales de la década de 1990, los investigadores en agentes se dieron cuenta de que, sin interfaces estándar, era difícil conseguir que agentes de diferentes fuentes trabajaran juntos. Esto cobra especial importancia cuando hablamos de sistemas multiagente abiertos en los que interactúan agentes de diferentes desarrolladores que pueden entrar y salir del sistema multiagente. En la actual proliferación de plataformas de agentes, también observamos fragmentación: una plataforma puede no comunicarse fácilmente con otra. La industria de la IA haría bien en recordar el valor de los estándares. Es alentador que estén surgiendo algunas iniciativas, como el Protocolo de Contexto de Modelos (MCP) de Anthropic en 2023 (Anthropic, 2023), cuyo objetivo es estandarizar la forma en que los asistentes de IA (agentes) interactúan con los sistemas externos. O A2A (Agent2Agent Protocol) (A2A Protocol, 2024), un estándar abierto que permite a los agentes de IA comunicarse y colaborar entre diferentes plataformas y marcos, independientemente de las tecnologías subyacentes. Se trata de un paso en la dirección correcta, pero se necesita más colaboración para evitar una situación en la que cada plataforma de agentes esté aislada de las demás.

Otro aspecto que se confunde en los medios es la equiparación excesiva del concepto de agente con los LLM. Si bien el entusiasmo actual se debe en gran medida a las capacidades de los LLM, el concepto de agente no exige el uso de un LLM. Un agente podría utilizar la lógica tradicional basada en reglas, una política de aprendizaje por refuerzo o cualquier combinación de técnicas de IA. De hecho, es probable que las mejores soluciones combinen varias técnicas de IA; por ejemplo, un enfoque híbrido de IA en el que el razonamiento simbólico (para la planificación a largo plazo o para garantizar las restricciones) complemente métodos subsimbólicos como las redes neuronales para la percepción o el procesamiento del lenguaje natural. Algunos han denominado a esta convergencia IA neurosimbólica. Las arquitecturas de agentes son un lugar natural para dicha convergencia: un agente podría utilizar un modelo neuronal para interpretar una situación y un planificador simbólico para decidir una secuencia de acciones, o un LLM para proporcionar capacidad de conversación al agente. Al reintegrar la investigación sobre agentes con la IA moderna, podemos lograr sistemas que realmente exhiban una inteligencia robusta.

8. HACIA LA INTEGRACIÓN: ABRAZANDO EL LEGADO DE LA INVESTIGACIÓN EN AGENTES EN LA IA MODERNA

¿Cómo podemos avanzar de manera que nos beneficiemos tanto de los nuevos avances como de los fundamentos establecidos? En primer lugar, la comunidad de IA (tanto investigadores como profesionales) debe esforzarse por utilizar una terminología clara y mantener la coherencia conceptual, siempre basada en el rigor científico. Si está creando un sistema en el que interactúan múltiples componentes basados en LLM, tenga en cuenta que ha creado un sistema multiagente y consulte la bibliografía sobre MAS para obtener orientación sobre el diseño y los posibles escollos. Si su aplicación implica un ciclo de toma de decisiones autónomo, considérela un agente inteligente y aproveche las arquitecturas de agentes conocidas (reactivas, BDI, etc.) para estructurarla. En publicaciones y debates, el uso de términos estándar (quizás con una nota que indique que se ajustan al término popular "Agentic AI") facilitará la interacción entre los investigadores experimentados en agentes y la nueva generación de desarrolladores entusiasmados con los agentes basados en LLM.

Utilizar el término IA agentiva (Agentic AI) no es incorrecto, si entendemos la IA agentiva como un estadio de integración neurosimbólica que articula el conocimiento estructurado y el razonamiento generativo, donde se integran los avances de la IA generativa en las teorías de agentes y sistemas multiagente preexistentes. El término sistemas multiagentivos (multiagentic systems) es, en mi opinión, más cuestionable y no tiene rigor científico; debería ser utilizado el término sistemas multiagente.

En segundo lugar, debemos incorporar activamente los resultados probados del área de los agentes en los nuevos marcos de agentes basados en LLM. Por ejemplo, el

concepto de un lenguaje de comunicación entre agentes podría ser muy relevante si empezamos a tener múltiples agentes basados en LLM que se comunican entre sí. Actualmente, la mayor parte de la comunicación entre agentes en los sistemas LLM se realiza en lenguaje natural ad hoc (literalmente, un agente da instrucciones a otro en lenguaje natural). Esto está bien para la creación de prototipos, pero a medida que aumenta la complejidad, podría redescubrirse la necesidad de un intercambio más estructurado (para evitar malentendidos y reducir la verbosidad). Los investigadores de la década de 1990 ya desarrollaron protocolos de comunicación basados en actos de habla; estos podrían inspirar esquemas de solicitud más eficientes o el paso de mensajes basados en API entre agentes. Trabajos recientes como el protocolo de comunicación autorreferencial (*Self-Referential Communication,* SRC) para agentes LLM (Barak, 2024) o Agora (*A Scalable Communication Protocol for Networks of Large Language Models*) (Marro *et al.,* 2024) se hacen eco de estas ideas, aunque expresadas en términos modernos.

Del mismo modo, las estrategias de coordinación y competencia de los MAS (como los algoritmos de asignación de tareas, los protocolos de consenso, los mecanismos de votación, la negociación automática, la argumentación automática, etc.) pueden transferirse a los sistemas de IA basados en agentes LLM para gestionar múltiples agentes que trabajan en subproblemas. Si, por ejemplo, se crea un equipo de agentes de IA basados en LLM para actuar como fuerza de ventas virtual (uno busca clientes potenciales, otro escribe correos electrónicos, otro programa llamadas), surge un problema de asignación y programación de recursos multiagente. En lugar de un enfoque ingenuo de prueba y error, se podrían aplicar los algoritmos de coordinación MAS existentes.

Pero no olvidemos que en muchos otros ámbitos problemáticos tendremos que desarrollar sistemas multiagente abiertos y, para ello, tendremos que tener en cuenta los resultados y conceptos ya desarrollados en el campo de los sistemas multiagente, principalmente las normas, en general las tecnologías de acuerdo, para desarrollar coreografías de agentes.

Por otro lado, la comunidad de investigación sobre agentes también debe actualizar sus herramientas aprovechando las ventajas de los LLM y el aprendizaje automático moderno. Las arquitecturas clásicas de agentes a menudo tenían dificultades con la adquisición de conocimientos y la flexibilidad, áreas en las que los LLM destacan. Un agente BDI con un planificador basado en LLM o un módulo de percepción basado en LLM podría ser mucho más potente que un agente BDI puramente simbólico. La integración del aprendizaje en las arquitecturas de los agentes (para que estos mejoren con la experiencia) ha sido un reto desde hace mucho tiempo. Las técnicas actuales de aprendizaje por refuerzo y aprendizaje con pocos ejemplos podrían finalmente hacerlo posible en la práctica. En resumen, la sinergia entre la sabiduría de los agentes tradicionales y las nuevas capacidades de la IA puede ampliar las fronteras de los sistemas autónomos.

9. CONCLUSIÓN

La "Agentic AI" se ha convertido en una palabra de moda que ha capturado la imaginación del mundo tecnológico, pero, como hemos argumentado, se trata esencialmente de un cambio de nombre de conceptos ya establecidos sobre agentes inteligentes autónomos, contextualizados por los recientes avances en IA generativa. El entusiasmo que suscitan los agentes autónomos basados en LLM está justificado por sus impresionantes nuevas capacidades, pero no debe llevarnos a ignorar el vasto conjunto de conocimientos sobre agentes y sistemas multiagente desarrollados a lo largo de décadas. El término "agentic", tomado de las ciencias sociales, destaca el comportamiento intencional y orientado a objetivos, precisamente el foco de la investigación sobre agentes inteligentes desde la década de 1990. Del mismo modo, los escenarios "multiagentic" son simplemente sistemas multiagente con otro nombre. Utilizar estos términos como si denotaran ideas novedosas corre el riesgo de generar confusión y repetir trabajos y errores del pasado.

Al revisar los fundamentos teóricos (la agencia de Bandura, los sistemas intencionales de Dennett) y los logros previos de la comunidad de IA (definiciones, propiedades, arquitecturas y plataformas para agentes autónomos), hacemos hincapié en que la rueda ya se ha inventado en lo que respecta a los agentes autónomos. Afortunadamente, es una rueda bastante buena, capaz de avanzar con el impulso de la IA generativa. Animamos a los profesionales actuales a basarse en investigaciones previas. En términos prácticos, esto implica adoptar la terminología correcta, consultar la bibliografía sobre sistemas de agentes/multiagentes para obtener orientación e integrar principios de diseño probados (como estándares de comunicación, arquitecturas de agentes, tecnologías de acuerdo, etc.) en los nuevos sistemas de IA.

El término "agentiva" proviene de la psicología social (Bandura, 1986) y denota la capacidad de actuar intencionalmente (Dennett, 1971). Su adopción en la IA moderna busca resaltar la supuesta intencionalidad de los agentes lingüísticos. Sin embargo, esta resemantización introduce redundancia conceptual, al atribuir "agencialidad" a entidades (agentes) que ya la poseen por definición. La llamada IA agentiva no introduce una nueva forma de agencia, sino un nuevo medio para expresarla lingüísticamente. El riesgo radica en la inflación terminológica: "agentes agentivos" o "IA multiagentiva" son tautologías que debilitan la precisión científica.

Por el contrario, también reconocemos que las nuevas tecnologías a menudo requieren adaptaciones de ideas antiguas. La comunidad de agentes debe aprovechar las oportunidades que ofrecen los LLM y el aprendizaje profundo para mejorar las capacidades de razonamiento y aprendizaje de los agentes. El resultado de combinar estos enfoques serán agentes autónomos verdaderamente avanzados, robustos, adaptables y capaces de abordar tareas complejas del mundo real. De hecho, como sugiere una visión, el futuro podría estar en los agentes neurosimbólicos, sistemas que combinan la percep-

ción y la comprensión del lenguaje impulsadas por la IA generativa con el razonamiento simbólico y la planificación. En mi opinión, esta integración de agentes, IA generativa, IA débil e IA responsable (RAI) es el camino hacia la IA neurosimbólica o IA híbrida, que es el siguiente paso en la evolución de la IA y también probablemente el futuro cercano de la IA "con cuerpo".

En conclusión, el campo de la IA debe tener cuidado de no perder de vista su propia historia en medio del entusiasmo actual. Los conceptos de agentes y sistemas multiagente siguen siendo tan relevantes en 2025 como lo fueron en 1995, incluso si los implementamos hoy con herramientas mucho más potentes. En lugar de acuñar términos ambiguos como "Agentic AI", deberíamos llamar a estos sistemas por su nombre y construir sobre los cimientos ya establecidos. Al hacerlo, no solo reconocemos el mérito donde corresponde, sino que también aceleramos el progreso, evitando volver sobre lo ya conocido y centrándonos en innovaciones genuinas. La rueda de los agentes inteligentes no necesita reinventarse; necesita perfeccionarse y potenciarse con los motores de la IA generativa. El camino que tienen por delante los agentes autónomos y sistemas multiagente es apasionante y, al unir los conocimientos del pasado con la tecnología actual, podemos garantizar que se construya sobre una ruta ya trazada, que nos lleve a destinos más allá de lo que las generaciones anteriores de IA fueron capaces de alcanzar.

La IA agentiva no representa una ruptura o innovación conceptual, sino una evolución tecnológica del paradigma de agentes autónomos y sistemas multiagente. Su relevancia reside en la capacidad de integrar los avances de la IA generativa dentro de marcos de acción autónoma preexistentes. La continuidad teórica entre los agentes clásicos y los sistemas basados en LLM es evidencia de la madurez del campo. El desafío actual no es definir nuevos tipos de agentes, sino preservar la coherencia teórica y evitar la fragmentación conceptual. Podemos ver la IA agentiva como un estadio de integración neurosimbólica que articula el conocimiento estructurado y el razonamiento generativo. Esta visión coincide con la necesidad, ya señalada por Wooldridge (2002) y Ferber (1999), de dotar a los sistemas multiagente de capacidades cognitivas más ricas sin perder su base formal. En consecuencia, la IA agentiva no sustituye a los agentes autónomos/SMA, sino que los reinterpreta como ecosistemas de agentes lingüísticos, donde la interacción se da en lenguaje natural y la coordinación se logra mediante mecanismos de argumentación y diálogo. El futuro de la disciplina dependerá de la capacidad para combinar la tradición formal de la teoría de agentes con las nuevas herramientas generativas y multimodales.

Referencias

Anthropic. (2023). Model Context Protocol (MCP) – Standard proposal by Anthropic for connecting AI assistants with external systems.

Austin, J. L., and Urmson, J. O. (1975). *How to do things with words* (2nd ed.). Cambridge, Mass.: Harvard University Press. ISBN 978-0674411524.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice- Hall. (Introduced the term "agentic" in context of human agency).

Bratman, M. E. (1999) [1987]. *Intention, Plans, and Practical Reason*. CSLI Publications. ISBN 1-57586-192-5.

Brooks, R. (1986). A Robust Layered Control System for a Mobile Robot. *IEEE Journal on Robotics and Automation, 2*(1): 14–23. (Describes the subsumption architecture, a reactive agent design).

Dennett, D. C. (1971). Intentional Systems. *Journal of Philosophy, 68*(2), 87–106. (Philosophical basis for attributing beliefs and desires to systems).

Ferber, J. (1999). *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence.* Addison-Wesley.

Finin, T., Fritzson, R., McKay, D., McEntire, R. (1994). KQML as an agent communication language. Proceedings of the third international conference on Information and knowledge management - CIKM 94. p. 456. doi:10.1145/191246.191322. ISBN 0897916743. S2CID 1129799. http://www.fipa.org/repository/aclspecs.html

FIPA. Foundation for Intelligent Physical Agents – (Standards organization for agent interoperability, est. 1996).

Gates, Bill. (2023). The Age of AI has Begun – GatesNotes Blog, Nov 9, 2023. (Discusses personal AI assistants and agents revolutionizing computing).

Genesereth, M. and Fikes, R. (June 1992). Knowledge Interchange Format Version 3.0 Reference Manual (PDF). Stanford Logic Group Report. Logic-92-1. Stanford University. Retrieved 7 August 2014. http://www.ksl.stanford.edu/software/ontolingua/

IBM – Think Blog. (2023). Agentic AI vs AI Agents: Understanding the Difference. IBM Think Topics. (Explains agentic AI as a framework vs AI agents as components).

IFAAMAS. International Foundation for Autonomous Agents and Multiagent Systems – http:// (Professional association supporting the AAMAS conference and MAS research).

Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems, 1*(1), 7–38. (Overview of the state of agent R&D in late 90s; covers architectures, applications, challenges).

Marro, S., La Malfa, E., Wright, J., Li, G., Shadbolt, N., Wooldridge, M., Torr, P. (2024). *A Scalable Communication Protocol for Networks of Large Language Models. arXiv:2410.11905.*

Ossowski, S., Sierra, C., & Botti, V. (2012, November 28). *Agreement Technologies: A Computing Perspective. Law*, Governance and Technology Series. Springer Netherlands. http://doi.org/10.1007/978-94-007-5583-3_1. https://dev.to/tomer_barak/self-reference-in-ai-agents-2nlh

Rao, A. S., and Georgeff, M. P. (1991). Modeling Rational Agents within a BDI-Architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, (473–484).

Rao, A. S., and Georgeff, M. P. (1995). BDI-agents: From Theory to Practice Archived 2011-06-04 at the Wayback Machine. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS"95), San Francisco, 1995*.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th Edition). Pearson. (AI textbook defining agents and environments; Chapter 2 on intelligent agents).

Searle, J. R. (1969). Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press (Verlag). ISBN 978-0-521-09626-3.

Shoham, y. (1993). Agent-oriented programming. *Artificial Intelligence, Volume 60,* Issue 1, 51-92. ISSN 0004-3702.

Sierra Garcia, C., Botti Navarro, V. J. Ossowski, D. S. (2011). Agreement Computing. *KI - Künstliche Intelligenz, 25*(1), 57-61. https://doi.org/10.1007/s13218-010-0070

Wooldridge, M. (2002). *An Introduction to Multi-Agent Systems.* John Wiley & Sons. (Textbook covering MAS definitions, architectures, and examples).

Wooldridge, M. (2009). *An Introduction to Multiagent Systems* (2nd ed.). Wiley.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152. (Classical paper defining intelligent agents and their properties).

CAPÍTULO V

Panarquía artificial: ciclos adaptativos de la IA

David Martinez Rego*

Este trabajo propone un marco analítico para comprender la evolución reciente de la inteligencia artificial (IA), superando la narrativa simplista de "veranos e inviernos". A través de la teoría de ciclos adaptativos y panarquía, se interpreta la IA como un sistema sociotécnico complejo, donde invención, innovación y difusión interactúan en múltiples escalas. Se revisan hitos históricos, desde la teoría estadística hasta el aprendizaje profundo y los grandes modelos de lenguaje, y se analizan limitaciones, anomalías empíricas y la transición hacia aplicaciones prácticas. El enfoque multinivel permite reconciliar expectativas extremas con la realidad, ofreciendo una visión prospectiva para la gestión estratégica de la tecnología.

Palabras clave: inteligencia artificial, panarquía, ciclos adaptativos, teoría del aprendizaje, aprendizaje profundo.

* Agradecemos a Funcas la invitación al evento y la oportunidad de participar en la presente publicación.

1. INTRODUCCIÓN

El debate contemporáneo en torno a la inteligencia artificial (IA) se caracteriza por una tensión constante entre entusiasmo desmedido y preocupación social, alimentada en gran medida por discursos mediáticos, intereses económicos y posicionamientos geopolíticos. En este contexto, resulta cada vez más difícil construir marcos analíticos que permitan comprender el estado real de la tecnología, sus límites estructurales, sus capacidades emergentes y su evolución futura desde una perspectiva científicamente fundamentada.

El objetivo de este trabajo es proponer un marco interpretativo para analizar la evolución reciente de la inteligencia artificial, alejándose del modelo narrativo dominante de "veranos e inviernos de la IA". Se argumenta que dicho modelo, aunque intuitivo, resulta insuficiente y potencialmente engañoso para comprender dinámicas complejas y acumulativas que caracterizan el desarrollo actual de la disciplina.

Como alternativa, se introduce un enfoque inspirado en la teoría de sistemas complejos, concretamente los conceptos de ciclo adaptativo y panarquía, originalmente desarrollados en el ámbito de la ecología por Holling y colaboradores (Holling, 2001). Este marco permite analizar la IA como un componente integrado en un ecosistema sociotécnico, sujeto a dinámicas de conservación, ruptura, reorganización y explotación que operan simultáneamente en múltiples escalas.

Es fundamental reconocer que el marco de comprensión que adoptemos condiciona no solo nuestras proyecciones mentales sobre el futuro de la inteligencia artificial, sino también nuestras decisiones colectivas frente a cambios tecnológicos. La teoría de sistemas enfatiza que un análisis incorrecto de un sistema complejo puede conducir a decisiones sociales y políticas indeseables, generando consecuencias contraproducentes a nivel económico, institucional o ético (Meadows and Wright, 2008). En el caso de la IA, la falta de un marco sólido de interpretación puede amplificar sesgos, malentendidos o expectativas erróneas, afectando la manera en que se implementan, regulan y adoptan estas tecnologías. Por ello, estudiar y formalizar un marco analítico riguroso para la IA no es solo un ejercicio académico, sino una necesidad estratégica para guiar la interacción de la sociedad con estas herramientas emergentes de manera informada, responsable y sostenible.

2. EL MODELO DE "VERANOS E INVIERNOS" DE LA IA

La narrativa de los "veranos e inviernos" ha sido ampliamente utilizada tanto en divulgación como en literatura académica para describir periodos alternantes de avances espectaculares y estancamiento de una tecnología. Según este modelo, las expectativas generadas por nuevos enfoques técnicos conducen a ciclos de inversión excesiva, seguidos inevitablemente por desilusión cuando dichas expectativas no se materializan.

2.1. Origen y difusión del modelo

Un ejemplo paradigmático de este ciclo corresponde al periodo comprendido entre finales de la década de 1980 y principios de los 2000. Hacia finales de los años 80 se produjo lo que podría considerarse un *verano* de la IA, caracterizado por un entusiasmo renovado en torno a los modelos neuronales y su potencial para la generalización, sustentado por avances en la retropropagación y en redes convolucionales (Rumelhart *et al.*, 1986; LeCun *et al.*, 1989). Sin embargo, la experiencia práctica pronto mostró que estos modelos carecían de la capacidad de generalizar de manera amplia y eficiente, limitando su aplicabilidad más allá de casos específicos. Esta situación llevó a un *invierno* prolongado, que se extendió aproximadamente hasta principios de la década del 2000, durante el cual la investigación en redes neuronales perdió prominencia frente a enfoques alternativos, como los sistemas expertos y los métodos de aprendizaje supervisado basados en teoría estadística (Mitchell, 1997). La reflexión crítica de muchos expertos en esta época es feroz y subrayaba que la imitación literal de procesos biológicos, aunque instructiva, no necesariamente conducía al desarrollo de sistemas de aprendizaje artificial efectivos[1]. Este periodo evidencia que las expectativas generadas durante los 80 con los modelos neuronales no estaban fundadas sobre ninguna base sólida, resaltando la importancia de comprender los límites teóricos y las condiciones de aplicabilidad de cada enfoque de aprendizaje.

Reacción a las limitaciones de los modelos neuronales

Las limitaciones prácticas encontradas por los modelos neuronales a finales de la década de 1980 impulsaron el surgimiento de avances conceptuales que hoy constituyen los fundamentos del aprendizaje automático moderno. Por un lado, la incapacidad de las redes neuronales de generalizar más allá de conjuntos de datos relativamente pequeños motivó el desarrollo de la teoría del aprendizaje estadístico. Por otro lado, la constatación de que ningún algoritmo de aprendizaje podría resolver de manera óptima todos los problemas posibles condujo a la formulación del teorema de *No Free Lunch* (NFL), que estableció límites universales a la optimización algorítmica (Wolpert and Macready, 1997).

Teoría del aprendizaje estadístico

El desarrollo del aprendizaje automático durante las décadas de 1980 y 1990 estuvo fuertemente influido por la teoría del aprendizaje estadístico, formalizada principalmente por Vapnik y Chervonenkis (Vapnik, 1998). Uno de sus resultados centrales establece que el error de generalización de un modelo depende de su complejidad (capacidad) y del tamaño del conjunto de datos de entrenamiento (ver figura 1).

[1] Vapnik (1998) llega a comparar la aproximación con inspirarse en el vuelo de las aves para la construcción de aviones.

FIGURA 1

PRINCIPIO FUNDAMENTAL DEL APRENDIZAJE ESTADÍSTICO ESTABLECE QUE EL ERROR DE GENERALIZACIÓN DE UN ALGORITMO A SOBRE UN PROBLEMA P ESTÁ ACOTADO SUPERIORMENTE POR LA SUMA DEL ERROR OBSERVADO EN EL CONJUNTO DE ENTRENAMIENTO D Y UN TÉRMINO PROPORCIONAL A LA RELACIÓN ENTRE LA COMPLEJIDAD DEL ALGORITMO A Y EL TAMAÑO DEL CONJUNTO D. DURANTE EL "INVIERNO" ORIGINAL DE LA IA EN LA DÉCADA DE 1990, LOS MODELOS NEURONALES FUERON CONSIDERADOS EXCESIVAMENTE COMPLEJOS EN RELACIÓN CON LOS CONJUNTOS DE DATOS DISPONIBLES, LO QUE SUGERÍA QUE SU ERROR EN LA PRÁCTICA NO PODÍA SER CONTROLADO, EXPLICANDO ASÍ SU LIMITADA ADOPCIÓN EN APLICACIONES REALES

$$\text{Error}(A,P) \leq \text{Error}_{\text{Entrenamiento}}(A,D) + \frac{\text{Complejidad}(A)}{\text{Tamaño}(D)}$$

De manera simplificada, puede afirmarse que modelos con mayor capacidad requieren conjuntos de datos proporcionalmente mayores para evitar el sobreajuste y un mal funcionamiento en la práctica. Este principio influyó profundamente en la percepción de las redes neuronales, consideradas excesivamente complejas para los recursos computacionales y los datos disponibles en ese momento.

El teorema de No Free Lunch

El teorema de *No Free Lunch* (NFL), formulado por Wolpert y Macready (1997), reforzó este escepticismo al demostrar que no existe un algoritmo de optimización universalmente superior cuando se promedian todos los posibles problemas. Este resultado fue interpretado como una limitación fundamental a la idea de algoritmos generales de aprendizaje, favoreciendo enfoques altamente especializados.

Efectos académicos y políticos del modelo

Estos dos desarrollos teóricos no solo fueron adoptados como explicación al estancamiento temporal de las redes neuronales durante la época, sino que ofrecían un certificado teórico de que la aproximación carecía de fundamento práctico, pues la cantidad de datos y cómputo necesarios para asegurar su funcionamiento práctico nunca iban a estar disponibles y su admisión como principio universal de aprendizaje contradice el principio de *No Free Lunch*.

Como consecuencia, tanto la academia como la industria aceptaron de manera generalizada el *statu quo* impuesto por las limitaciones de los modelos neuronales y redujeron drásticamente la financiación destinada a esta línea de investigación. La mayoría

de los investigadores en redes neuronales abandonaron progresivamente este enfoque, desplazándose hacia el desarrollo de modelos especializados de aprendizaje supervisado que pudieran garantizar resultados prácticos en aplicaciones concretas (Mitchell, 1997). Este cambio de trayectoria consolidó un paradigma centrado en modelos de capacidad limitada y conjuntos de datos relativamente pequeños, fomentando la investigación aplicada en problemas específicos y ralentizando el progreso en el desarrollo de arquitecturas neuronales.

3. EL RESURGIMIENTO EMPÍRICO DEL APRENDIZAJE PROFUNDO

El resurgimiento del aprendizaje profundo en la última década se ha manifestado de manera particularmente visible a través de dos fenómenos tecnológicos y sociales: la eficacia de los modelos de procesamiento de imágenes y la irrupción de los grandes modelos de lenguaje (más conocidos por la siglas de su nombre anglosajón *Large Language Models* o los LLM) en aplicaciones de interacción natural con la información textual. El punto de inflexión definitivo se da en el año 2022, cuando el uso masivo de ChatGPT da visibilidad generalizada de los LLM en la sociedad.

Desde la perspectiva tradicional de los ciclos de "veranos e inviernos" de la IA, este fenómeno se ha interpretado de dos maneras contrapuestas. Por un lado, existe una corriente que mantiene que la teoría del aprendizaje estadístico sigue siendo válida y robusta, y que los avances actuales no hacen sino evidenciar los límites que tarde o temprano aparecerán nuevamente, anticipando un futuro "invierno" de la IA. Por otro lado, hay quienes consideran que los inviernos históricos han quedado atrás y que los principios teóricos eran, en el mejor de los casos, irrelevantes o incluso incorrectos para comprender la dinámica de los modelos neuronales modernos.

Como alternativa a estas interpretaciones simplistas, sostenemos que este hito no representa un cambio abrupto en la teoría, sino el resultado de un proceso acumulativo de más de dos décadas. Durante este tiempo, los principios de diseño derivados de la teoría estadística han permanecido presentes, guiando tanto la construcción de modelos como la organización de conjuntos de datos y el desarrollo de la industria de procesamiento de información. La expansión de tecnologías para la captura y procesamiento masivo de datos permitió la observación de fenómenos empíricos que antes eran inaccesibles, revelando límites prácticos de la teoría y comportamientos de aprendizaje que, aunque no anticipados, resultan reproducibles en distintos dominios y tareas. Estos resultados empíricos, obtenidos bajo condiciones de datos masivos y cómputo distribuido, permiten extender la teoría del aprendizaje estadístico: muestran cómo las estructuras internas de los modelos neuronales pueden alinearse con tareas generales, producir generalización efectiva y aprender representaciones útiles sin supervisión estricta, construyendo un marco empírico que amplía y refina los fundamentos teóricos originales.

A pesar de la complejidad inherente a la explicación de la evolución de un campo académico dinámico y multidimensional como el aprendizaje automático, resulta útil, a efectos expositivos, identificar fases generales que permiten comprender cómo se ha configurado el camino hacia las tecnologías actuales de aprendizaje profundo y grandes modelos de lenguaje.

3.1. 2000-2010: auge de la ciencia de datos y los modelos especializados

Entre finales de los años noventa y principios de la década de 2010, emergió un paradigma pragmático que hoy se reconoce como Ciencia de Datos. Este enfoque priorizaba la construcción de modelos matemáticos específicos para cada problema, con estructuras relativamente simples, optimización convexa y propiedades estadísticas bien comprendidas. El dominio de esta aproximación se alinea con una mayoría académica que promulga la adherencia a principios teóricos como camino al éxito. Aunque los modelos de aprendizaje automático de la época tenían una precisión limitada debido a su simplicidad matemática —lo que permitía que la teoría explicara su comportamiento y generara expectativas predecibles—, esta misma simplicidad facilitó la creación de modelos predictivos razonablemente eficaces a partir de los conjuntos de datos relativamente pequeños de la época. Esta capacidad fue suficiente para habilitar modelos de negocio dominantes en la actualidad, especialmente aquellos basados en recomendación, personalización y predicción, que sustentaron de manera decisiva el crecimiento y consolidación de empresas como Google, Amazon o Netflix. En este sentido, la aparente limitación impuesta por la adherencia a la teoría de los modelos no impidió su impacto práctico y económico; su predictibilidad y manejabilidad contribuyeron a transformar la industria de la información y sentaron las bases de los sistemas de datos masivos.

Sin embargo, hacia el final de este período, este enfoque comenzó a evidenciar límites claros y estructurales que condicionaron su potencial de innovación. En primer lugar, la necesidad de diseñar modelos matemáticos robustos y *ad hoc* para cada problema específico implicaba costos elevados de desarrollo, tanto en términos de tiempo de investigación como de recursos humanos y computacionales, generando fricciones significativas que restringían la velocidad y amplitud de la innovación. En segundo lugar, se observaron rendimientos decrecientes en términos de precisión: los primeros modelos representaban avances sustanciales respecto a enfoques previos, pero las iteraciones posteriores, aunque refinadas, no lograban incrementos relevantes en la exactitud ni en la utilidad práctica de los sistemas, proporcionando únicamente mejoras marginales que difícilmente justificaban nuevas inversiones significativas. Finalmente, se identificaron límites claros en la escalabilidad: cuando estos modelos cuidadosamente diseñados dentro de los marcos teóricos existentes se aplicaban a dominios de entrada más complejos, como imágenes de alta resolución o grandes volúmenes de texto, su desempeño se degradaba de manera notable y las arquitecturas se rompían, revelando que las soluciones exitosas en problemas controlados no podían generalizar de manera eficiente a escenarios más amplios y

realistas. Estos factores combinados explican por qué el enfoque ad hoc y altamente matemático llega a un estancamiento.

3.2. 2010-2016: *big data*

La transición en la década de 2010 estuvo marcada por un cambio de escala tanto en los datos como en la infraestructura de procesamiento disponible para la investigación y la aplicación del aprendizaje automático. Esta etapa, conocida comúnmente como *big data*, puede entenderse como una evolución natural del paradigma de aprendizaje supervisado basado en la teoría estadística desarrollado entre 2000 y 2010. Durante la fase anterior, los modelos matemáticos específicos para cada problema habían demostrado ser efectivos y relativamente predecibles, permitiendo desarrollar sistemas prácticos de recomendación, predicción y personalización. Sin embargo, estos modelos alcanzaron rápidamente límites estructurales: requerían costos elevados de desarrollo, presentaban rendimientos decrecientes en precisión y problemas de escalabilidad a dominios más complejos, como imágenes de alta resolución o grandes volúmenes de texto.

El fenómeno *big data* surgió como una respuesta técnica y económica a estos límites. Por un lado, la expansión de la captura de datos y la disponibilidad de grandes volúmenes de información permitieron plantear modelos de aprendizaje más ambiciosos, que podrían aprovechar conjuntos de datos mucho mayores que los que habían sido posibles hasta la fecha. Esta expansión no solo incluyó el aumento de la cantidad de datos generados por usuarios, sensores, transacciones comerciales, redes sociales, registros biomédicos y dispositivos conectados, sino también un incremento en la diversidad de aplicaciones.

Para gestionar esta enorme cantidad de datos, se desarrollaron tecnologías específicas de almacenamiento y procesamiento distribuido. Entre las más relevantes destacan los sistemas de archivos distribuidos como Hadoop Distributed File System (HDFS), que permiten almacenar *petabytes* de información de manera redundante y escalable, y *frameworks* de procesamiento distribuido como Hadoop MapReduce y Spark, que facilitan la ejecución de operaciones de transformación y agregación de datos de forma paralela en clústeres de servidores. Estas herramientas, combinadas con el surgimiento de soluciones de almacenamiento en la nube (*cloud computing*) y redes de alta velocidad, crearon un ecosistema capaz de soportar el manejo masivo de información, democratizando el acceso a capacidades que anteriormente solo podían ser asumidas por grandes empresas con infraestructura propia.

Este incremento en la disponibilidad de datos y recursos de procesamiento generó, a su vez, un cambio cultural en la investigación y la industria. La posibilidad de trabajar con conjuntos de datos masivos abrió nuevas oportunidades para el aprendizaje supervisado, pero también evidenció que los enfoques tradicionales basados en aprendizaje estadístico tenían limitaciones para justificar el retorno de la inversión (ROI) de esta

infraestructura. Modelos relativamente simples y *ad hoc*, diseñados para problemas específicos, ya no podían aprovechar plenamente la escala de los datos ni los recursos computacionales disponibles. La industria y la academia comenzaron a percibir que, para aprovechar estos conjuntos de datos y justificar las inversiones en infraestructura, era necesario explorar aproximaciones más generales y empíricas que pudieran capturar patrones complejos sin requerir un diseño ad hoc para cada caso.

De este modo, la fase *big data* constituye un puente crítico hacia el resurgimiento del aprendizaje profundo. Por un lado, consolidó los principios heredados de la teoría de aprendizaje estadístico (Vapnik, 1998; Shalev-Shwartz, 2014): la necesidad de modelos robustos, la relación entre complejidad y tamaño de datos y la importancia de buenas técnicas de optimización como guía teórica. Por otro lado, generó un entorno propicio para experimentación empírica masiva, donde los límites de la teoría se podían observar de manera práctica y reproducible gracias a la abundancia de datos y capacidad de cómputo. Esta combinación de teoría sólida, infraestructura escalable y *datasets* masivos sentó las bases para el desarrollo de arquitecturas neuronales profundas más grandes y complejas, capaces de abordar dominios como visión por computadora y lenguaje natural, que posteriormente conducirían al surgimiento de modelos de lenguaje a gran escala (LLM) y al resurgimiento empírico del aprendizaje profundo.

3.3. 2008-2016: Mafia de la IA y anomalías

Mientras la Ciencia de Datos y el paradigma *big data* se consolidaban en el mercado a partir de técnicas de aprendizaje estadístico bien fundamentadas teóricamente, se desarrollaba en paralelo una línea de investigación menos dominante pero persistente centrada en redes neuronales profundas. Esta corriente, a menudo asociada a un núcleo reducido de investigadores —frecuentemente representados por Geoffrey Hinton, Yann LeCun y Yoshua Bengio, conocidos coloquialmente como la "Mafia de la IA" y galardonados con premios como el Nobel y la Medalla Turing—, aprovechó de manera sistemática el aumento progresivo de la capacidad computacional, el abaratamiento del almacenamiento y la disponibilidad de grandes volúmenes de datos para extender el estudio empírico del aprendizaje profundo (LeCun *et al.*, 2015; Bengio *et al.*, 2013). A diferencia del enfoque estadístico clásico, estas investigaciones exploraron la combinación de técnicas supervisadas y no supervisadas con el objetivo de aprender representaciones internas jerárquicas y reutilizables, capaces de capturar regularidades latentes en los datos. El resultado fundamental de estos esfuerzos fue la constatación empírica de que los modelos de aprendizaje profundo podían superar de forma consistente la precisión de los enfoques estadísticos tradicionales en dominios donde estos últimos habían alcanzado un estancamiento prolongado, como el reconocimiento de imágenes, el procesamiento del habla y, posteriormente, el lenguaje natural (Hinton *et al.*, 2006). Este avance no supuso una negación explícita de los principios del aprendizaje estadístico, sino una extensión práctica de sus límites observables, demostrando que bajo condiciones adecuadas de

datos y cómputo era posible aprender representaciones robustas que transformaban problemas previamente intratables en tareas abordables mediante modelos entrenables de extremo a extremo.

Cabe destacar que, durante este período, comenzaron a acumularse una serie de observaciones empíricas clave que resultaban difíciles de ignorar desde el marco teórico dominante en ese momento. Estos resultados, reproducidos de manera consistente en distintos laboratorios y dominios de aplicación, se manifestaban como anomalías empíricas que no podían explicarse adecuadamente a partir de la teoría del aprendizaje estadístico clásica ni de sus supuestos sobre complejidad, generalización y escalabilidad. Lejos de tratarse de excepciones aisladas, estas observaciones señalaron patrones sistemáticos en el comportamiento de los modelos neuronales profundos bajo regímenes de datos y cómputo masivos, dando lugar a una nueva tendencia de investigación centrada en comprender y explotar estos fenómenos. En las secciones siguientes se detallan dichas observaciones empíricas y se analiza su papel en la transformación del campo del aprendizaje automático.

3.3.1. Aprendizaje de representaciones mediante observación no supervisada

Una de las primeras observaciones empíricas que contribuyó de manera decisiva al resurgimiento del aprendizaje profundo fue la constatación de que es posible aprender estructuras internas útiles para la resolución de problemas complejos mediante técnicas no supervisadas, es decir, a partir de la mera observación de los datos sin necesidad de anotaciones humanas explícitas. Este hallazgo contradecía una de las limitaciones prácticas más relevantes de los enfoques dominantes hasta ese momento: la fuerte dependencia de grandes volúmenes de datos etiquetados, cuyo coste de obtención constituía un cuello de botella tanto en investigación como en aplicaciones industriales.

El trabajo seminal de Hinton y Salakhutdinov (Hinton, 2006) demostró que redes neuronales profundas, entrenadas de manera no supervisada para reducir la dimensionalidad de los datos, eran capaces de aprender representaciones latentes que capturaban regularidades estructurales relevantes del dominio subyacente. Estas representaciones no solo preservaban información esencial, sino que facilitaban de forma significativa el posterior aprendizaje supervisado, permitiendo alcanzar mejores resultados con una fracción de los ejemplos anotados previamente necesarios.

Desde una perspectiva sistémica, este enfoque introdujo un cambio cualitativo en la manera de diseñar algoritmos de aprendizaje. En lugar de construir modelos específicos para cada tarea desde cero, se hizo posible reutilizar modelos preentrenados como componentes genéricos de representación, adaptables a múltiples problemas mediante un ajuste fino posterior. Esta reutilización redujo de manera drástica el esfuerzo de ingeniería algorítmica, disminuyó la dependencia de datos etiquetados y sentó las bases concep-

tuales de los actuales modelos fundacionales, marcando un punto de inflexión respecto a los paradigmas de aprendizaje supervisado especializados que habían dominado la disciplina hasta entonces.

3.3.2. Escalado de datos y arquitecturas profundas en visión artificial

Una segunda observación empírica fundamental emergió a partir de la convergencia entre la disponibilidad de conjuntos de datos de una escala sin precedentes y la aplicación de arquitecturas neuronales profundas entrenadas de manera eficiente. El proyecto ImageNet constituyó, en este sentido, la mayor base de datos de imágenes naturales creada hasta la fecha con fines de investigación, proporcionando millones de ejemplos etiquetados que cubrían una amplia diversidad de objetos y contextos visuales. Este recurso transformó el estudio de la visión artificial al situar el problema en un régimen de datos radicalmente distinto al previamente explorado.

Sobre esta base empírica, el trabajo de Krizhevsky *et al.* (2012) demostró que una arquitectura neuronal conocida —las redes convolucionales profundas, previamente exploradas sin éxito concluyente en décadas anteriores—, combinada con innovaciones clave en los procedimientos de entrenamiento y el uso intensivo de *hardware* especializado, podía superar de forma contundente a todos los algoritmos de visión artificial desarrollados en los treinta años previos. El modelo conocido como AlexNet redujo de manera drástica el error en la competición ImageNet, marcando una ruptura clara con el estado del arte dominado hasta entonces por métodos basados en características diseñadas manualmente y clasificadores estadísticos especializados.

Este resultado tuvo un impacto profundo en la comunidad científica al evidenciar que los límites observados en los enfoques estadísticos clásicos no respondían necesariamente a barreras fundamentales del problema, sino a restricciones impuestas por la escala de los datos, la capacidad de los modelos y los métodos de optimización disponibles. La combinación de grandes volúmenes de datos, arquitecturas profundas y capacidad de cómputo suficiente reveló un nuevo régimen empírico en el que la precisión y la capacidad de generalización continuaban mejorando de forma sistemática, desafiando las expectativas establecidas por la teoría dominante y consolidando al aprendizaje profundo como la aproximación central en visión por computador.

3.3.3. Arquitecturas profundas como generadores de contenido

Una de las contribuciones más significativas del aprendizaje profundo en este tiempo fue demostrar que los modelos neuronales no solo podían aprender representaciones útiles para clasificación o predicción, sino también generar contenido de manera autónoma y creativa, guiada por objetivos definidos por el usuario. Este avance se materializó con las

redes generativas antagónicas (GANs), introducidas por Goodfellow *et al.* (2014), que combinan un generador y un discriminador en un esquema de entrenamiento competitivo. Mediante este mecanismo, los modelos pueden producir ejemplos sintéticos que imitan las propiedades estadísticas del conjunto de datos de entrenamiento, pero además permiten al usuario dirigir la generación hacia características deseadas, ya sea en imágenes, audio u otros tipos de datos.

El impacto de este hallazgo es doble. Por un lado, confirma que las redes profundas aprenden representaciones internas suficientemente ricas como para producir contenido coherente y novedoso, incluso en dominios donde los métodos estadísticos clásicos no podían generar resultados útiles sin supervisión intensiva. Por otro lado, abre un nuevo paradigma de aplicaciones prácticas: la generación creativa dirigida por el usuario, que incluye desde síntesis de imágenes artísticas hasta la creación de datos sintéticos para entrenamiento y validación de otros modelos, o interfaces interactivas donde el aprendizaje profundo responde a instrucciones humanas. En consecuencia, esta observación ilustra cómo el aprendizaje profundo no solo amplía los límites de precisión y generalización, sino que también redefine lo que significa "aprender" y "crear" dentro de un marco computacional, consolidando una nueva categoría de anomalías empíricas que desafían la teoría estadística tradicional.

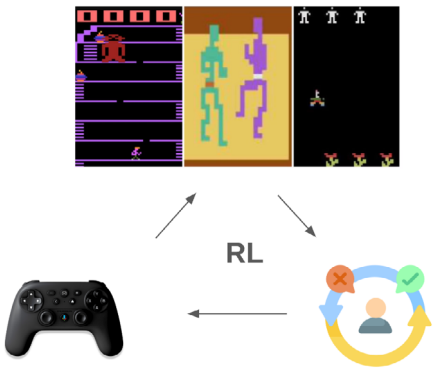
3.3.4. Redes neuronales profundas, aprendizaje por refuerzo y simuladores

Una observación empírica crítica durante la consolidación del aprendizaje profundo fue la demostración de que, dado un sistema de control general, un simulador suficientemente realista del sistema a controlar y un criterio de retroalimentación adecuado, es posible entrenar modelos de aprendizaje por refuerzo profundo para generar soluciones de control efectivas sin intervención humana directa. El trabajo seminal de Mnih *et al.* (2015) ilustró este principio mediante la aplicación de redes neuronales profundas a entornos de videojuegos complejos, mostrando que, con entrenamiento suficiente en episodios simulados y utilizando una señal de recompensa bien diseñada, los agentes podían aprender políticas de control que superaban el rendimiento humano en varias tareas.

El elemento clave de este avance radica en la combinación de tres factores: primero, la disponibilidad de un simulador que reproduzca de manera realista la dinámica del sistema, permitiendo la acumulación de experiencia en millones de episodios sin riesgos físicos; segundo, la formulación de un criterio de retroalimentación que guíe el aprendizaje hacia comportamientos deseables; y tercero, la capacidad de la red neuronal profunda para construir representaciones internas ricas y jerárquicas del problema, que facilitan la planificación de acciones complejas y creativas. Esta arquitectura permitió que los modelos no solo aprendieran soluciones de control óptimas, sino que también desarrollaran secuencias de acciones novedosas y efectivas, aplicables en entornos del mundo real.

FIGURA 2

EL APRENDIZAJE POR REFUERZO COMBINADO CON SIMULADORES Y UNA FUNCIÓN OBJETIVO BIEN DEFINIDA CONSTITUYEN EL MARCO DE SISTEMAS DE CONTROL MODERNOS



Fuente: Volodymyr *et al.* (2015).

En conjunto, este hallazgo evidencia que los sistemas de aprendizaje profundo, cuando se combinan con simulación de alta fidelidad y señales de retroalimentación adecuadas, pueden transformar problemas de control general en tareas abordables mediante aprendizaje autónomo. Esto abre la puerta a aplicaciones prácticas en robótica, vehículos autónomos y sistemas de automatización industrial, demostrando que el aprendizaje por refuerzo profundo no solo extiende las capacidades predictivas de las redes neuronales, sino que también permite generar comportamiento creativo y adaptativo en dominios complejos y dinámicos.

3.3.5. Ciencia de Datos 2.0

El período comprendido entre 2016 y 2025 puede caracterizarse como una fase emergente de la disciplina que hemos denominado Ciencia de Datos 2.0. Esta etapa se fundamenta en la convergencia de anomalías empíricas observadas en los años anteriores y la disponibilidad de recursos tecnológicos sin precedentes. Las observaciones previas sobre la capacidad de los modelos neuronales profundos para aprender representaciones útiles, generar contenido creativo y controlar sistemas complejos mediante aprendizaje por refuerzo, dieron lugar a una nueva tendencia: diseñar arquitecturas neuronales profundas robustas, incrementar su tamaño y complejidad, y combinarlas con conjuntos de datos masivos y heterogéneos. La práctica habitual de preentrenamiento no supervisado, seguida de optimización supervisada o reforzada, permite aprovechar patrones subyacentes en los datos y facilita la generalización, reproduciendo de manera consistente fenómenos que antes eran considerados anomalías (Doshi *et al.*, 2019; Hooker *et al.*, 2019).

Entre estas anomalías destaca el fenómeno conocido como *grokking*, en el que modelos suficientemente complejos, expuestos a grandes cantidades de datos y entrenados durante tiempos prolongados con esquemas de regularización adecuados, aprenden de manera estructural la solución subyacente a un problema, separando claramente memorizar de generalizar. La reproducibilidad de este tipo de eventos permite avanzar en la comprensión teórica de los límites del aprendizaje profundo y extender la teoría estadística clásica, integrando observaciones empíricas que previamente se consideraban excepciones (Doshi *et al.,* 2019).

El éxito de la Ciencia de Datos 2.0 no puede separarse de la coincidencia con avances en *hardware*, fenómeno denominado *hardware lottery* (Hooker, 2021). La disponibilidad de GPUs de alta eficiencia, almacenamiento masivo y arquitectura distribuida permitió que estas redes complejas pudieran entrenarse en tiempos razonables, haciendo visibles los efectos observables y reproducibles que definieron la tendencia. Es importante notar que los modelos actuales no constituyen la única vía posible para desarrollar tecnologías de este tipo; son, más bien, aquellos que pueden implementarse eficazmente con el *hardware* existente, condicionando la evolución empírica y la adopción industrial.

El ejemplo más público de Ciencia de Datos 2.0 es ChatGPT, donde un protocolo de aprendizaje por refuerzo orientado a "jugar el juego del lenguaje" fue expuesto al mundo, acumulando millones de episodios lingüísticos y permitiendo que un modelo de lenguaje a gran escala aprenda una política efectiva de interacción con el mundo (Long *et al.*, 2022). En este caso, los principios algorítmicos originales aplicados a videojuegos, como los agentes de Atari, se mantuvieron intactos; únicamente el controlador del juego fue sustituido por tokens de lenguaje, y el juego consiste en interaccionar de manera coherente y productiva en un entorno lingüístico abierto. Este enfoque demuestra que la combinación de arquitecturas profundas, grandes conjuntos de datos de entrenamiento y algoritmo de retropropagación del error pueden producir sistemas capaces de comportamientos complejos, creativos y útiles, consolidando Ciencia de Datos 2.0 como la etapa contemporánea dominante en inteligencia artificial.

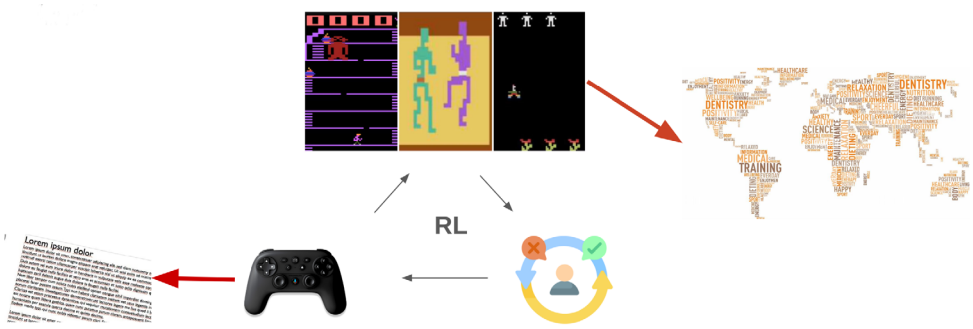
4. PANARQUÍA ARTIFICIAL

Tras la revisión de los últimos veinticinco años de avances en inteligencia artificial y aprendizaje estadístico, resulta evidente que lo que a primera vista puede interpretarse como un fracaso de una tecnología y un éxito de otra constituye, en realidad, un conjunto de avances incrementales que son teóricamente coherentes y complementarios en la práctica. La teoría y la experiencia acumuladas muestran que existen numerosos problemas que aún se resuelven de manera más eficiente mediante enfoques de aprendizaje estadístico, lo que explica que la industria asociada a estos métodos continúe activa y próspera.

Visualizar el panorama de la IA en términos de tecnologías ganadoras o perdedoras, como propone la narrativa de "veranos e inviernos", conduce con frecuencia a decisiones equivocadas en múltiples niveles, desde la inversión hasta la formulación de políticas públicas. Si bien es cierto que, como hemos detallado, la aparición de resultados empíricos significativos puede generar cambios de perspectiva y reorganización en el campo, esto no implica que los avances previos hayan perdido completamente su valor ni que los nuevos desarrollos reemplacen de manera inmediata las dinámicas existentes. Por ello, para evitar juicios sesgados derivados de un enfoque estrictamente estacional, resulta necesario adoptar un marco de análisis que considere la interacción entre sociedad, tecnología y acontecimientos científicos. Dicho marco permitiría evaluar de manera integral el estado de las tecnologías disponibles en cada momento, reconociendo simultáneamente la continuidad de los conocimientos previos, la emergencia de innovaciones y la influencia de factores socioeconómicos en la adopción y desarrollo de estas. Una aproximación de este tipo posibilitaría decisiones más informadas, equilibradas y estratégicas en la planificación de investigación, inversión y políticas públicas en el ámbito de la inteligencia artificial.

FIGURA 3

LOS MODELOS DE LENGUAJE ACTUALES PROVIENEN DE UN USO EFECTIVO DE TÉCNICAS DE APRENDIZAJE POR REFUERZO EN EL CASO DEL LENGUAJE NATURAL



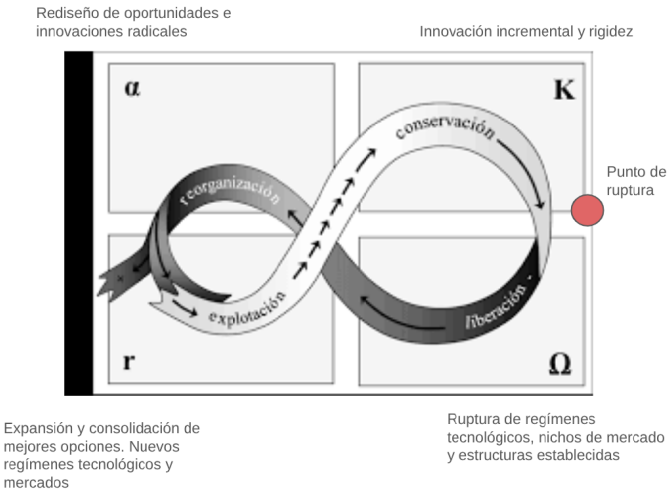
5. PANARQUÍA Y CICLOS ADAPTATIVOS EN INTELIGENCIA ARTIFICIAL

El concepto de *panarquía* y los *ciclos adaptativos* provienen de la teoría de sistemas complejos y fueron desarrollados inicialmente en el contexto de la ecología por Holling y colaboradores (Holling, 2001). Su objetivo principal es describir cómo los sistemas complejos, ya sean naturales, sociales o tecnológicos, evolucionan a través de fases recurrentes de estabilidad, colapso y reorganización. La intuición básica es que lo que puede parecer un comportamiento caótico o impredecible a nivel de un sistema específico puede, de hecho, estar estructurado por ciclos de adaptación que operan en distintas escalas, interactuando entre sí.

En general, un *ciclo adaptativo* puede entenderse como un proceso de cuatro fases: *conservación*, *liberación*, *reorganización* y *explotación*. Estas fases evolucionan de manera continua e indefinida, tal y como representa la figura 4. En la fase de *conservación*, los recursos, conocimientos y estructuras del sistema se consolidan, aumentando su eficiencia y especialización, pero al mismo tiempo incrementando su rigidez frente a perturbaciones externas. Posteriormente, la fase de *liberación* ocurre cuando una perturbación significativa, interna o externa, rompe la estabilidad del sistema, liberando los recursos acumulados y permitiendo que las restricciones anteriores desaparezcan. Esta liberación da paso a la fase de *reorganización*, en la que los elementos del sistema se reconfiguran, se prueban nuevas interacciones y emergen nuevas estrategias de funcionamiento. Finalmente, durante la fase de *explotación*, las innovaciones y reconfiguraciones exitosas se consolidan, los recursos se utilizan de manera eficiente, y el sistema entra nuevamente en un período de conservación, cerrando el ciclo. Cada fase es esencial: la conservación de recursos y conocimiento permite estabilidad y eficiencia, mientras que la liberación y reorganización facilitan la innovación y la adaptabilidad.

FIGURA 4

VISUALIZACIÓN DE UN CICLO ADAPTATIVO



El modelo de ciclos adaptativos ha sido utilizado con éxito para explicar fenómenos complejos en contextos muy diversos, desde ecosistemas forestales y pesqueros (Holling, 1995), hasta sistemas socioeconómicos como mercados financieros y organizaciones urbanas (Gunderson and Holling, 2002). En todos estos casos, los cambios que a primera vista parecen abruptos o desordenados pueden interpretarse de manera más precisa como transiciones entre ciclos recurrentes de conservación y reorganización.

Aplicando esta perspectiva al campo de la inteligencia artificial, podemos reinterpretar la narrativa de los "veranos e inviernos" como manifestaciones superficiales de ciclos adaptativos.

Durante reorientaciones fundamentales en la IA como campo de conocimiento, como los finales de los años 80 con los modelos neuronales iniciales o los recientes resurgimientos del aprendizaje profundo y los modelos de lenguaje a gran escala, el sistema científico pasa por fases de reorganización y explotación iniciadas por una experiencia que rompe con *statu quo* científico (conservación) anterior.

Las invenciones generadas se estudian de manera intensiva, los recursos (datos, computación y capital humano) se concentran y la eficiencia de los sistemas recientemente ideados aumenta. Sin embargo, estas fases también van generando rigidez: limitaciones teóricas, cuellos de botella computacionales y dificultades de generalización, etc. que eventualmente conducen a una nueva fase de liberación. Durante esta nueva fase de reorganización el ciclo no se frena sino que se repite. Investigadores y empresas exploran nuevas arquitecturas, combinan técnicas supervisadas y no supervisadas, desarrollan infraestructuras de *Big Data* o experimentan con simuladores y aprendizaje por refuerzo para redefinir lo que es posible. Así emergen los avances que posteriormente entrarán en explotación: redes neuronales profundas escaladas, modelos generativos como GANs, y más recientemente, modelos de lenguaje a gran escala que interactúan con el mundo a través de retroalimentación humana (como son ChatGPT y otros sistemas comerciales similares). Es importante destacar que cada ciclo no elimina ni rompe completamente el valor de los desarrollos anteriores. Por ejemplo, los métodos estadísticos y los modelos especializados continúan siendo hoy útiles en dominios donde su eficiencia e interpretabilidad son superiores y comparten raíces teóricas con los modelos neuronales.

Si consideramos el avance en IA como una serie de ciclos adaptativos y aspiramos a tener un modelo completo, no podemos ignorar la percepción de "veranos e inviernos", sino que necesitamos ampliar el modelo de manera que esta percepción sea explicable. Para ello es necesario reconocer y modelar que los ciclos de invención en IA no ocurren en un sistema aislado, sino que influencian y son influenciados por otros ciclos adaptativos con escalas espaciales y temporales que pueden ser diferentes. Este es el rol del concepto de *panarquía*, el cual añade una dimensión adicional al permitir observar múltiples ciclos adaptativos interactuando a distintas escalas. El observar múltiples ciclos adaptativos de manera conectada ayuda a reconciliar situaciones paradójicas como narrativas de "invierno" acerca de tecnologías que están avanzando o "verano" acerca de tecnologías cercanas a su límite de rendimiento. En la siguiente sección se introduce nuestro modelo de panarquía para el caso de la IA.

TABLA 1

PRINCIPALES FASES EN EL RESURGIMIENTO DEL APRENDIZAJE PROFUNDO

	Fase	**Ciencia de Datos**
2000-2010	Observaciones científicas	Modelos especializados simples, enfoque *ad hoc*, optimización convexa, propiedades estadísticas bien comprendidas
	Impacto práctico	Sistemas de recomendación y personalización; consolidación de empresas como Google, Amazon y Netflix
	Herramientas/Técnicas	Modelos lineales y bayesianos, *datasets* pequeños
	Fase	***Big Data***
2010-2016	Observaciones científicas	Disponibilidad masiva de datos y cómputo; límites estructurales de modelos estadísticos evidenciados
	Impacto práctico	Infraestructura distribuida y democratización del procesamiento masivo; puente hacia arquitecturas neuronales profundas
	Herramientas/Técnicas	Hadoop, Spark, HDFS, *cloud computing*; aprendizaje supervisado escalable
	Fase	**Mafia de la IA y anomalías**
2008-2016	Observaciones científicas	Redes neuronales profundas aprenden representaciones jerárquicas; anomalías reproducibles; superan modelos estadísticos en visión, lenguaje y habla
	Impacto práctico	Éxitos en visión por computador (ImageNet), generación de contenido (GANs), aprendizaje por refuerzo; reducción de dependencia de datos etiquetados
	Herramientas/Técnicas	Redes convolucionales profundas, preentrenamiento no supervisado, GANs, simuladores para RL
	Fase	**Ciencia de Datos 2.0**
2016-2025	Observaciones científicas	Consolidación de anomalías empíricas; preentrenamiento seguido de ajuste fino; fenómenos como *grokking*
	Impacto práctico	Modelos fundacionales escalables (LLMs, ChatGPT); aplicaciones en lenguaje, control y generación de contenido
	Herramientas/Técnicas	Arquitecturas profundas masivas, GPUs, cómputo distribuido, aprendizaje por refuerzo en lenguaje natural

5.1. Panarquía artificial

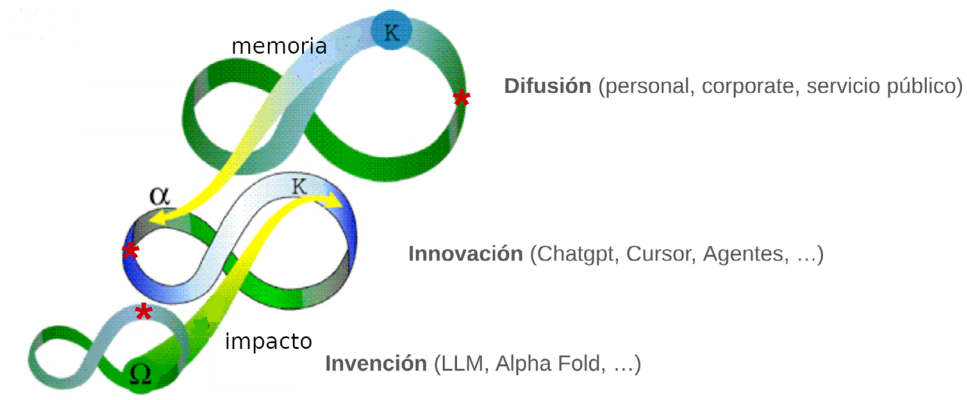
La entrada de los modelos de lenguaje a gran escala (LLM) en la esfera pública ha generado una disonancia notable entre expectativas, regulación y realidad. Por un lado, se han propagado afirmaciones sobre la cercanía de una superinteligencia de uso general

(*Artificial General Intelligence* o AGI) con propuestas de regulaciones orientadas a la no proliferación similares al caso de la tecnología nuclear bélica. Además, se han movilizado niveles de inversión sin precedentes para financiar su desarrollo y despliegue. Sin embargo, la adopción de la tecnología, varios años después de su irrupción inicial, no ha generado los cambios radicales esperados; los retornos sobre el capital invertido siguen siendo inciertos, y aunque su uso es amplio, se limita a aplicaciones muy específicas. La coexistencia de estas realidades ha generado una polarización hacia dos extremos, uno de confianza ciega en la proximidad de una AGI y otro de espera por el colapso inevitable de la tecnología.

El análisis de esta situación compleja sin recurrir a percepciones requiere un marco analítico más completo, capaz de reconciliar las dinámicas de invención, innovación y consolidación tecnológica. Proponemos interpretar la situación a través de un modelo de *panarquía multinivel*, compuesto por ciclos adaptativos interconectados a diferentes escalas, tal y como se representa en la figura 5. La división en niveles propuesta se inspira en la clasificación sugerida por Narayanan y Kapoor (2025):

FIGURA 5

PANARQUÍA ARTIFICIAL Y ESTADO A FECHA DE PUBLICACIÓN DE ESTE LIBRO



1. *Ciclo de difusión*: abarca la adopción de la tecnología en la vida cotidiana, en diversos ámbitos como el personal, el corporativo y los servicios públicos. Este nivel refleja la integración progresiva de la IA en rutinas establecidas, regulaciones y estructuras sociales existentes, donde la inercia humana y los marcos regulatorios actúan como filtros que ralentizan la adopción, pero también protegen contra riesgos potenciales.

2. *Ciclo de innovación:* corresponde a la transformación de la tecnología en productos utilizables y servicios concretos, como ChatGPT, Cursor u otras aplicaciones comerciales basadas en LLM. Aunque los ciclos de innovación pueden ser más ágiles que los de difusión, es importante tener en cuenta que la implementación de productos innovadores conlleva tiempos e inercias, incluso cuando se utilizan tecnologías base plenamente funcionales.

3. *Ciclo de invención*: refleja los ciclos generados por descubrimientos en laboratorios de investigación o empresas específicas, donde surgen nuevas tecnologías fundacionales, como los propios LLM.

Desde esta perspectiva multinivel, resulta más sencillo comprender la situación actual. A nivel de invención, los ciclos recientes han reforzado la supremacía del aprendizaje profundo, limitando temporalmente la exploración de caminos alternativos. Los avances son incrementales y se concentran en perfeccionar las líneas de investigación dominantes, lo que genera rigidez tecnológica y focaliza los recursos —tanto de capital como humanos— en la expansión y consolidación de esta tecnología.

A nivel de innovación, la tecnología se ha encontrado con limitaciones: a) para trasladar sus capacidades a aplicaciones prácticas concretas y b) para alcanzar niveles de rendimiento suficientes que permitan su uso en problemas de mayor impacto, como son las tecnologías de agentes. Aunque el avance hacia casos de uso de mayor impacto ha sido más lento que el ciclo de invención, el ciclo de innovación está superando su fase de reorganización. La mayoría de productos digitales de consumo han reevaluado su posicionamiento a la luz de las nuevas capacidades y se vislumbra una fase de explotación, en la que la tecnología puede alcanzar una penetración masiva y consolidar las mejores prácticas de diseño y despliegue.

Finalmente, en el ciclo de difusión, la adopción está modulada por regulaciones, inercia institucional y estructuras sociales consolidadas, que retrasan la penetración completa de la tecnología, pero al mismo tiempo actúan como un mecanismo de mitigación de riesgos.

Mientras la narrativa de "veranos e inviernos" promueve un enfoque reactivo, donde las decisiones se toman en función de percepciones de éxito o fracaso a corto plazo y en una sola escala temporal, la perspectiva de *panarquía* permite un manejo más estratégico de la evolución tecnológica. Al analizar los distintos niveles de invención, innovación y difusión en relación con la fase del ciclo adaptativo en la que se encuentran, es posible diseñar intervenciones y estrategias mejor contextualizadas. Por ejemplo, las regulaciones existentes, las prácticas institucionales consolidadas y la experiencia adquirida durante difusiones tecnológicas previas ya actúan como mecanismos de memoria (*remember*) que orientan la implantación en IA, evitando riesgos innecesarios y promoviendo desarrollos más sostenibles. De manera complementaria, una vez que la fase de conservación actual en la investigación en aprendizaje profundo llegue a su límite, se espera que nuevas observaciones disruptivas den lugar a nuevos ciclos de investigación,

utilizando los recursos, conocimiento y aprendizajes acumulados. La visión de panarquía no solo explica la coexistencia de estabilidad social, innovación paulatina e invención disruptiva en el ecosistema de la inteligencia artificial, sino que también proporciona un marco operativo para gestionar su evolución de manera proactiva, reconociendo la interacción de diferentes ciclos a diferentes escalas.

5.1.1. Panarquía artificial actual

Una manera de poner a prueba el marco de panarquía consiste en analizar de manera prospectiva la situación actual de la inteligencia artificial, considerando la interacción de ciclos adaptativos a distintos niveles.

En el nivel de invención, es de esperar que la fase de conservación actual se vea interrumpida por algún avance súbito o por un límite significativo. Entre los factores que podrían provocar esta ruptura se incluyen:

- Modelos pequeños y manejables que progresivamente demuestran capacidades cercanas a modelos propietarios de gran escala, cuestionando la necesidad del consumo de grandes recursos de computación.

- Dificultades para encontrar fuentes de energía o infraestructura de *hardware* que soporten modelos más grandes, generando cuellos de botella que podrían ralentizar la evolución.

Paralelamente, la investigación en esta fase se ha basado en *benchmarks* académicos que en un primer momento, parecían resolver problemas prácticos deseados. Sin embargo, pronto se observó que los avances académicos no se traducían en productos útiles, lo que motivó el desarrollo de *benchmarks* más realistas y orientados a aplicaciones prácticas. Este desfase entre los objetivos de la investigación académica y las expectativas de innovación explica en gran parte el avance lento en la difusión de la tecnología percibido recientemente. Esta brecha ha comenzado a cerrarse progresivamente, aunque todavía persisten limitaciones significativas en muchos casos de uso deseados.

En el ciclo de innovación, se ha superado la fase de reorganización, en la que se han detectado limitaciones concretas de la tecnología a la hora de crear aplicaciones prácticas. Esta fase de experimentación, aunque puede interpretarse de manera pesimista, ha generado guías de diseño más robustas para sistemas prácticos basados en IA, preparando el terreno para una fase de explotación. Entre los ejemplos recientes de limitaciones catalogadas destacan:

- Limitaciones en la capacidad de razonamiento frente a problemas complejos (Shojaee *et al.,* 2025).

- Dificultades en sistemas de frontera para mantener consistencia y coherencia en tareas de razonamiento general (Kamradt *et al.,* 2025).

- Pérdida de coherencia en conversaciones multiturno de modelos de lenguaje (Zhou *et al*., 2025).

- Restricciones en el manejo de contextos extensos, limitando la aplicabilidad práctica de LLMs en tareas complejas (Kriman *et al.*, 2025).

Por otra parte, las expectativas financieras y valoraciones bursátiles de la innovación en IA pueden actuar como disruptor negativo en el ciclo de innovación actual. Una corrección significativa de expectativas de retorno de inversión podría conducir a un abandono masivo de la tecnología.

Finalmente, en el ciclo de difusión, existen limitaciones naturales que han retrasado la adopción masiva de la IA en la vida pública:

- No todo el conocimiento humano está formalizado o escrito; gran parte es tácito y social, dificultando su uso por sistemas de IA.

- Rigidez en organizaciones, procesos de decisión y estructuras institucionales que retrasan la integración de nuevas tecnologías.

- Automatización parcial: solo alrededor del 10 % de muchos procesos puede ser efectivamente automatizado con IA.

- Problemas de gestión del cambio y fricciones regulatorias, tanto sociales como externas (por ejemplo, limitaciones energéticas).

Estas barreras explican por qué la disrupción social esperada aún no ha ocurrido de manera masiva. Sin embargo, desde la perspectiva de los ciclos adaptativos, es previsible que la tecnología eventualmente se incorporará de manera más profunda en la sociedad. La interacción entre ciclos de invención, innovación y difusión indica que, aunque la adopción plena está siendo gradual, el avance se está produciendo en todos los niveles de la panarquía planteada.

6. CONCLUSIONES

El análisis de la inteligencia artificial mediante el marco de panarquía y ciclos adaptativos nos permite comprender de manera más matizada la dinámica tecnológica actual. La narrativa de "veranos e inviernos" es insuficiente para capturar la complejidad de la interacción entre investigación, innovación y adopción social.

Los avances disruptivos son resultado de una acumulación de observaciones de base científica que discuten una teoría dominante hasta reorientar nuestra comprensión de un campo de conocimiento e iniciar una reorganización de su estudio; el ciclo científico no avanza de manera aislada sino que interactúa con ciclos de innovación que generan productos con alto impacto en la sociedad; finalmente los ciclos de difusión, evidenciados a través de barreras institucionales, bloqueos comerciales e inercias sociales modulan la penetración de una tecnología, con consecuencias tanto positivas como negativas. Todo análisis o regulación novedosa en ámbitos sin precedentes —como la GDPR en el caso de la expansión del intercambio de datos o el AI Act en el caso de la IA—incorpora, de forma explícita o implícita, un modelo de evolución futura. Cuando dicho modelo no refleja de manera realista y exhaustiva el fenómeno que se pretende regular, existe el riesgo de desaprovechar oportunidades o de generar dinámicas no deseadas.

La perspectiva multinivel presentada ofrece una base sólida para analizar y anticipar la evolución de la IA a corto y medio plazo, integrar aprendizajes previos y gestionar riesgos y oportunidades de manera estratégica y contextualizada.

Referencias

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8),1798–1828.

Doshi,D., Das A., He, T., and Gromov, A. (2019). To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. *arXiv preprint arXiv:1912.*12349.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27.

Gunderson, L. H., and Holling, C. S. (2002). *Panarchy: Understanding Transformations in Human and Natural Systems*. Island Press.

Hinton, G. E., Osindero, S., and Yee-Whye Teh, Y-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7),1527–1554.

Holling, C. S. (1995). What barriers? what bridges? *Ecological Applications*, *5*(2), 523–525.

Holling, C. S. (2001). Understanding the complexity of economic, ecological, and social systems. *Ecosystems*.

Hooker, S. (2021). The hardware lottery. *Communications of the ACM, 64*(12), 56–63.

Hooker, S., Courville, A., Clark, G., and Dauphin, Y. (2019). What do compressed deep neural networks forget? *arXiv preprint arXiv:1910.* 01742.

Kamradt, G., Landers, B., Pinkard, H., Chollet, F., and Knoop, M. (2025). Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2501.00003*.

Kriman, S., Acharya, S., Hsieh, C. P., Sun, S. (2025). Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2501.*00005.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation, 1*(4), 541–551.

Meadows, D. H., and Wright, D. (2008). *Thinking in Systems: A Primer*. Chelsea Green Publishing.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mnih, V. *et al*. (2015). Human-level control through deep reinforcement learning. *Nature*.

Narayanan, A., and Kapoor, S. (2025). *Ai as normal technology: An alternative to the vision of AI as a potential superintelligence*. Knight First Amendment Institute.

Ouyang, L. *et al.* (2022). Training language models to follow instructions with human feedback. *arXiv*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536.

Shalev-Shwartz, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Shojaee, P. *et al.* (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2501.*00002.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Wolpert, D., and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation.*

Zhou, Y., Neville, J., Laban, P., and Hayashi, H. (2025). Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2501.*00004.

CAPÍTULO VI

Autonomía basada en valores: hacia una gobernanza responsable de los agentes IA

Holger Billhardt*
Alberto Fernández
Andrés Holgado
Sascha Ossowski

La inteligencia artificial a menudo se concibe como el proyecto de diseñar agentes artificiales capaces de tomar decisiones de forma autónoma y actuar de forma eficiente en entornos cada vez más complejos. Cuanto más sofisticadas se hagan las tareas que los humanos deleguemos en dichos agentes, menos podremos supervisar sus decisiones y mayor será el impacto (positivo o negativo) que tendrán sus acciones en nuestras vidas. Hay una concienciación creciente de que la autonomía de los sistemas IA ha de desempeñarse de forma responsable y, en particular, que sus decisiones y acciones han de estar alineadas con los valores (éticos) de la sociedad. En este capítulo se van a repasar algunas técnicas para gobernar los sistemas multiagente, y presentar modelos recientes para que los agentes IA puedan desempeñar su autonomía de forma responsable, alineando sus acciones con los valores éticos de la sociedad.

Palabras clave: inteligencia artificial ética, gobernanza de los sistemas multiagente, agentes conscientes de los valores.

* Este trabajo cuenta con el apoyo de las subvenciones: VAE: TED2021-131295B-C33, financiada por MCIN/AEI/10.13039/501100011033 y del programa "European Union NextGenerationEU/PRTR"; COSASS: PID2021-123673OB-C32, financiada por MCIN/AEI/10.13039/501100011033; EVASAI: PID2024-158227NB-C32, financiada por MICIU/AEI/10.13039/501100011033/FEDER, UE; y la beca "Contratos Predoctorales de Personal Investigador en Formación en Departamentos de la Universidad Rey Juan Carlos (C1 PREDOC 2025)", financiada por la Universidad Rey Juan Carlos para Andrés Holgado-Sánchez.

1. INTRODUCCIÓN

Estamos delegando cada vez más tareas en máquinas como si fuesen personas. Estas tareas pueden ser tan sencillas como pasar la aspiradora a un piso o más complejas como conducir un coche de forma autónoma. También interactuamos cada vez más con las máquinas como si fuesen personas. El ejemplo prototípico que conocemos todos es ChatGPT, pero hay muchas más. Por ejemplo, el registro en muchos hoteles en Asia ha de hacerse necesariamente a través de un sistema informático, y a veces este está complementado por un robot humanoide. Muchas veces, para realizar estas tareas, las máquinas tienen que interactuar, o bien cooperando, o bien compitiendo entre ellas. Por ejemplo, unos drones han de cooperar entre ellos para poder vigilar un área de forma exhaustiva; o los sistemas compiten entre ellos cuando participan en una subasta electrónica en línea. Las aplicaciones de estas características se llaman *sistemas multiagente*. Si dichos sistemas comprenden tanto entes computacionales como humanos, a menudo se habla también de *sistemas sociotécnicos* (Wooldridge, 2009).

Los entes computacionales que componen este tipo de aplicaciones a menudo se denominan *agentes* IA. Dichos agentes pueden ser o bien solo *software* o bien ciberfísicos, es decir, que pueden incluir un componente de *hardware* como los robots. En ambos casos, las personas pueden delegar tareas en los agentes IA, que estos realizan en representación de sus usuarios. Para tal fin necesitan un conjunto de capacidades indispensables: primero, han de ser *racionales*, es decir, que deben elegir la manera más eficiente para cumplir la misión que les ha sido encargada por sus usuarios. Segundo, han de ser *sociales*, es decir, tienen que ser capaces de interactuar con otros agentes, bien cooperando o bien compitiendo, para así poder cumplir una misión con éxito. Pero, sobre todo, tienen que ser *autónomos*. Autonomía aquí se entiende como la capacidad de cumplir con misiones sin la necesidad de una intervención humana directa y permanente. Por ejemplo, un granjero puede estar interesado en encargar a un conjunto de drones que fumiguen su campo, sin la necesidad de tener que indicarles que cada vez que se acerquen al límite del campo han de dar la vuelta. O, si uno de estos drones falla, los demás han de ser capaces de reorganizarse y acabar la tarea de forma transparente para el usuario.

Para actuar con autonomía, un agente IA afronta una serie de retos. Quizá la problemática más espinosa deriva del hecho de que el significado y el alcance de su autonomía han de ser *interpretados* adecuadamente en una multitud de situaciones que inicialmente son desconocidas. En primer lugar, hay que especificar muy bien el alcance de las misiones que un agente IA debe realizar y cuáles son exactamente las *metas* que conseguir, a sabiendas de que hay mucha incertidumbre respecto a los entornos y los contextos en los que el agente va a actuar. Esto es una instancia de un problema recurrente en inteligencia artificial llamado frecuentemente "Problema del rey Midas", y que es especialmente relevante en el contexto de los agentes IA[1]. Por otro lado, hay que

[1] El legendario rey Midas, de la mitología griega, pidió que todo lo que tocara se convirtiera en oro, pero se arrepintió después de tocar comida, bebida y a sus familiares...

especificar exactamente cuáles son los *medios* que un agente IA puede emplear para cumplir la misión que le ha sido encargada. No todo vale para conseguir un objetivo, pero estas limitaciones a menudo no se hacen explícitas al definir la funcionalidad de un agente IA. Hay varios ejemplos en los que sistemas como ChatGPT proporcionan respuestas poco éticas, como por ejemplo acceder a la consulta de un niño y dar consejos de cómo proceder para suicidarse[2].

Un segundo reto importante es el hecho de que la autonomía tiene que ser adecuadamente *gobernada*. Un sistema multiagente normalmente constituye un sistema abierto, es decir, que de antemano es desconocido cuántos agentes van a formar parte del sistema, qué características tendrán y cómo se comportarán. Este es el caso en una subasta electrónica *online*, ya que no está claro cuántos agentes IA van a pujar por un bien a subastar, ni qué estrategia va a emplear cada uno de ellos. Cada agente IA va a actuar de forma individualmente racional, haciendo las pujas que piensa que más le convienen, y a menudo sin considerar los efectos de las acciones a nivel del sistema. Aun así, como diseñador del sistema global, nos gustaría que este tenga ciertas propiedades deseables, como, por ejemplo, eficiencia, equidad, robustez, etcétera. En nuestro ejemplo, el diseñador de un mercado electrónico podría imponer determinados protocolos de subasta, para asegurar que, aun siendo individualmente racionales y "egoístas", está en el interés de cada uno de los agentes no actuar de forma estratégica y revelar su valoración real del bien a subastar en sus pujas (Sandholm, 1999). Normalmente, en el área de los sistemas multiagente, esto se consigue mediante el diseño del *entorno* en el que están actuando los agentes IA, por ejemplo, la plataforma de *software* en la que los agentes IA han de registrarse para poder pujar por un bien a subastar (Schumacher and Ossowski, 2006). Otro buen ejemplo son las ciudades inteligentes, o espacios inteligentes en general, y que han sido posibles gracias al uso de las nuevas infraestructuras de comunicación ubicuas. Por un lado, dichas infraestructuras pueden dotar a los actores de nuevas facilidades y así *facilitar* nuevas formas de interacción. Por ejemplo, en el caso de los sistemas de transporte inteligentes, surge la opción de realizar comunicaciones entre vehículos (*vehicle to vehicle*, V2V) o entre vehículos y la infraestructura de gestión de tráfico (*vehicle to infrastructure*, V2I). Pero también permiten establecer dinámicamente qué actuaciones están permitidas o prohibidas y así *gobernar* las interacciones entre los agentes IA, es decir, influir en sus decisiones sin anular su autonomía.

El resto del artículo se organiza como sigue. La sección 2 se dedica al último reto mencionado, discutiendo métodos y mecanismos para gobernar sistemas multiagente abiertos, e ilustrando el problema a través del ejemplo de unos *gestores de intersección* autónomos en el ámbito de una gestión inteligente del tráfico. En la sección 3 se analiza el primero de los retos planteados. Se argumenta la necesidad de diseñar agentes IA cuyo comportamiento esté alineado con los valores de los usuarios a los que representan o de la sociedad en la que el correspondiente sistema sociotécnico está embebido. Se ilustra a través de la noción de los agentes IA conscientes de los valores, y un ejem-

[2] https://edition.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit

plo de su aplicación en el contexto del aprendizaje de sistemas de valores. La sección 4 resume la argumentación del artículo y apunta a retos abiertos y trabajos futuros.

2. GOBERNAR LOS SISTEMAS MULTIAGENTE

En esta sección se plantean retos y propuestas para la gobernanza de sistemas que involucran agentes IA autónomos. El apartado 2.1 las describe desde un punto de vista técnico, mientras que el apartado 2.2 las analiza desde una perspectiva regulatoria. El apartado 2.3 ilustra la argumentación mediante un ejemplo.

2.1. Protocolos para gobernar sistemas multiagente abiertos

En los sistemas multiagente abiertos, el diseñador no tiene un control absoluto sobre los agentes que forman el sistema ni sobre su comportamiento. Para poder alcanzar aún así una coordinación adecuada entre los agentes IA, habitualmente se diseñan las características del *entorno* en el que los agentes interactúan. Una forma común es mediante los *protocolos de interacción,* a través de los cuales los agentes se comunican entre sí. Por ejemplo, es conocido que en subastas de vuelta única, si los postores han de comportarse de acuerdo con un protocolo tipo Vickrey, la mejor estrategia para cada uno de ellos es revelar sus preferencias de forma certera (Vickrey, 1961; Sandholm, 1999). Nótese que la forma de analizar protocolos en este contexto es diferente a la que se aplica normalmente en informática: en los sistemas distribuidos comúnmente hay un interés en probar que todas las posibles trazas de ejecución permitidas por un protocolo cumplan ciertas propiedades como seguridad y vivacidad; en el caso de los protocolos de interacción de los sistemas multiagente, en cambio, se consideran normalmente solo en aquellas posibles trazas de ejecución en las que los agentes IA se comportan de forma individualmente racional, es decir, que toman las acciones que previsiblemente más les acercan al cumplimento de los objetivos que les han sido encargados por sus usuarios humanos.

Como se mencionó anteriormente, las ciudades inteligentes son un dominio de aplicación muy adecuado para los sistemas multiagente. Suele haber un gran número de ciudadanos que utiliza infraestructuras, como las de energía o de transporte, y cada vez más este uso está mediado a través de la tecnología. Esto es posible porque se han desarrollado plataformas de comunicación que permiten a los usuarios (y/o a los agentes IA que les representan) comunicarse entre sí y con los elementos de la infraestructura. En lo que sigue, se van a presentar algunos ejemplos de cómo los protocolos de interacción que ofrecen dichas plataformas de comunicación pueden, por un lado, *facilitar* nuevas formas de coordinación y, por otro, *gobernar* las interacciones entre los agentes.

Vasirani y Ossowski (2013) proponen un mecanismo de coordinación para *balancear* la carga de vehículos eléctricos en viviendas, en lugar de cargarlos todos simultáneamente, reduciendo de esta manera el impacto en la red de distribución. Se asume que la subestación eléctrica puede asignar energía de forma dinámica a cada puesto de recarga en diferentes instantes de tiempo (*time-slots*). El algoritmo de asignación de energía a los agentes se basa en una planificación por lotería (*lottery scheduling*), en la que los agentes tienen un número de boletos que pueden utilizar para acceder a los recursos, en este caso, energía eléctrica. De forma resumida, el protocolo seguido para asignar energía en un espacio de tiempo es el siguiente: (1) la subestación notifica a todos los agentes el número total de participantes (vehículos); (2) cada agente envía a la subestación el número de boletos que quiere jugar en ese reparto; (3) la subestación calcula el reparto en base a la cantidad de boletos emitidos por los agentes y activa cada punto de recarga con la energía asignada; (4) la subestación informa a cada agente un "tipo de cambio" que refleja el valor de los boletos que todavía tiene ese agente. Los autores demuestran de forma analítica que, en situaciones ideales con información completa, unas actuaciones individualmente racionales de los agentes llevan a un reparto de energía eficaz y que, mediante unos parámetros del protocolo, es posible influir en este "equilibrio eficiente" para que promueva más la eficiencia o la equidad. Mediante simulaciones validan que, en situaciones más realistas, los agentes pueden *aprender* este mismo equilibrio eficaz si existe un sistema normativo de penalizaciones adecuado.

Otro ejemplo se presenta en Dötterl *et al.* (2020), en el que se propone un sistema de transporte colaborativo en el que las personas (a través de sus agentes IA) se ofrecen a realizar pequeñas desviaciones en sus rutas de transporte previstas (por ejemplo, volviendo a casa después de trabajar) para transportar objetos de un punto a otro. Es interesante aquí destacar el protocolo seguido cuando un agente estima que va a entregar con retraso un paquete. En esas situaciones, un agente IA coordinador facilita la búsqueda de otro agente cercano que pueda entregarlo a tiempo. De manera general, (1) los agentes repartidores (disponibles o efectivamente transportando un paquete) transmiten periódicamente su información local (localización, dirección, medio de transporte usado, etc.); (2) los agentes repartidores que detectan que van con retraso solicitan al coordinador la transferencia de su paquete a otro agente; (3) el coordinador recibe la solicitud de transferencia, busca un potencial repartidor cercano y le ofrece la tarea; (4) si el agente candidato muestra interés, se propone la sugerencia de transferencia al agente que la solicitó. Si se llega al acuerdo, se propone la transferencia. En otro caso, se repiten los pasos (3) y (4), o se intenta más tarde. Mediante simulaciones, los autores demuestran la eficacia del mecanismo, aun cuando los agentes tomen decisiones guiados únicamente por su interés propio.

2.2. Sistemas multiagente abiertos y la regulación legal

La gobernanza de sistemas involucrando agentes IA autónomos no solo ha de plantearse desde un punto de vista técnico, sino que ha de considerar también su vertiente legal. En

particular, las soluciones propuestas a nivel técnico muchas veces representan pruebas de concepto que no abarcan o consideran todos los aspectos normativos (ni los éticos) que se deben tener en cuenta para su implantación real. Una posible implantación, por tanto, requiere un análisis profundo de la legalidad de las propuestas planteadas con el fin de identificar cambios que sean necesarios tanto a nivel técnico como a nivel legislativo para asegurar su conformidad con las normas legales.

En Santos *et al. (*2020) se propone un método para el análisis jurídico y la subsiguiente adaptación de propuestas técnicas para este tipo de sistemas. Con tal fin, se plantea una colaboración entre expertos técnicos y jurídicos que se estructura en tres fases principales.

La primera fase se centra en la especificación del funcionamiento del sistema. Es necesario identificar y especificar todos los aspectos del sistema que podrían afectar o verse afectados por cuestiones legales. Por un lado, dicha especificación debe incluir el dominio de aplicación a alto nivel del sistema, así como todas las entidades afectadas. Y, por otro lado, debe especificar las operaciones principales (que puedan afectar a la normativa legal), ocultando los detalles técnicos. El objetivo aquí consiste en identificar "qué" hace el sistema, no "cómo" lo hace, de tal forma que expertos jurídicos puedan analizar el cumplimiento de las normas vigentes.

La segunda fase analiza el cumplimiento legal. Los expertos jurídicos deben analizar si los diferentes aspectos identificados en el paso anterior se ajustan a la normativa actual. Este paso a menudo no es sencillo y puede requerir un análisis más detallado de diferentes interpretaciones de las normas existentes. Si se detecta una infracción, se deben identificar las normas o principios legales vulnerados y asociarlos al aspecto correspondiente del sistema.

En la última fase se acuerdan propuestas de modificación. Basándose en el análisis y la discusión del paso segundo, se deben identificar posibles soluciones que hagan que el sistema propuesto cumpla con la normativa. Dichas soluciones pueden consistir en propuestas de modificación de las normativas existentes, modificaciones que deban introducirse en el sistema propuesto o soluciones híbridas. Aquí se deben ponderar los efectos o consecuencias de modificar cada parte (normativa o sistema) para decidir cuál debe adaptarse. Por ejemplo, en el caso de detectarse un incumplimiento de derechos fundamentales, adaptar la normativa puede ser inviable, por lo que la modificación del sistema sería la única opción para garantizar el cumplimiento.

2.3. Un ejemplo: gestión de intersecciones basada en subastas

Vamos a ilustrar los retos para la gobernanza de los sistemas multiagente, tanto desde el punto de vista técnico como el legal y ético, a través de un ejemplo. Con tal fin,

nos centramos en una posible infraestructura de tráfico para un hipotético futuro con coches autónomos guiados por agentes IA.

Eliminar al conductor humano del bucle de control mediante el uso de vehículos autónomos integrados con una infraestructura vial inteligente puede considerarse como el objetivo último y a largo plazo del conjunto de sistemas y tecnologías agrupados bajo el nombre de Sistemas de Transporte Inteligente (*Intelligent Transportation Systems*, ITS). Las ventajas de dicha integración abarcan desde una mayor seguridad vial hasta un uso más eficiente de la red de transporte. Por ejemplo, los vehículos pueden intercambiar información crítica de seguridad con la infraestructura para reconocer situaciones de alto riesgo con antelación y, por lo tanto, alertar a los conductores. Además, los sistemas de semáforos pueden comunicar información sobre las fases y el tiempo de señalización a los vehículos para optimizar el uso de la red de transporte.

2.3.1. Gestión de intersecciones con vehículos autónomos

En este sentido, muchos autores han prestado recientemente atención al potencial de una integración más estrecha de los vehículos autónomos con la infraestructura para la gestión de intersecciones (Namazi *et al.,* 2019). Dresner y Stone (2008) introducen el sistema de control basado en reservas, donde una intersección es regulada por un componente de *software* inteligente, denominado gestor de intersección, que asigna reservas de espacio y tiempo a cada vehículo autónomo que pretende cruzar la intersección. Cada vehículo es operado por otro agente de *software*, llamado agente conductor. Cuando un vehículo se aproxima a una intersección, el agente conductor solicita al gestor de intersección reservar los espacios e intervalos de tiempo necesarios para cruzar la intersección de manera segura. Este último simula la trayectoria del vehículo dentro de la intersección e informa al agente conductor si su solicitud entra en conflicto con las reservas ya confirmadas de otros agentes. Si no existe tal conflicto, el agente conductor almacena los detalles de la reserva e intenta cumplirlos; de lo contrario, puede realizar una nueva solicitud más tarde. Los autores muestran mediante simulaciones que, en situaciones de tráfico equilibrado, si todos los vehículos son autónomos, sus retrasos en la intersección se reducen drásticamente en comparación con los semáforos tradicionales. Vasirani y Ossowski (2012) amplían el modelo anterior para el control de intersecciones basado en reservas en dos líneas principales. En primer lugar, para intersecciones individuales, elaboran una política basada en subastas para la asignación de reservas a los vehículos que concede acceso preferente según las diferentes preferencias de los conductores respecto a sus tiempos de viaje. En segundo lugar, para redes de intersecciones, establecen un mercado computacional donde los gestores de intersección compiten entre sí como proveedores de reservas y adaptan de manera egoísta los precios de reserva de las subastas que ejecutan, con el fin de ajustarse a la demanda real. De esta manera, los flujos de tráfico se distribuyen mejor en la red vial y se reduce el coste social de aplicar una estrategia de asignación preferente con respecto a la intersección. A continuación, se resumen los aspectos técnicos más relevantes de este enfoque, para seguidamente analizar las cuestiones éticas y legales asociadas.

2.3.2.Gestor de intersección basado en subastas

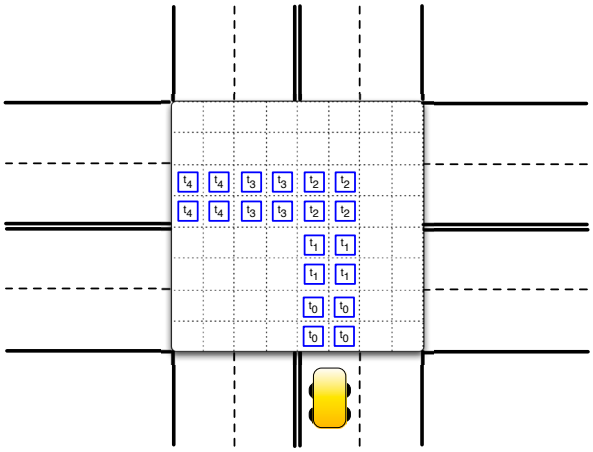
Para una única intersección basada en reservas, el problema que debe resolver un gestor de intersección consiste en asignar reservas entre un conjunto de conductores de manera que se maximice un objetivo específico. Este objetivo puede ser, por ejemplo, minimizar el retraso promedio causado por la presencia de la intersección regulada. En este caso, la política más simple a adoptar es asignar una reserva al primer agente que la solicite, como ocurre con la política de "primero en llegar, primero en ser atendido" (*first come, first served*, FCFS) propuesta por Dresner y Stone en su trabajo original (Dresner y Stone, 2008). Otro trabajo alineado con este objetivo se inspira en la teoría de colas adversarias para definir varias políticas de control alternativas que buscan minimizar el retraso promedio (Vasirani and Ossowski, 2021).

Sin embargo, estas políticas ignoran el hecho de que, en el mundo real, dependiendo del contexto y de su situación personal, las personas valoran de manera muy diferente la importancia de los tiempos de viaje y los retrasos. Dado que procesar las solicitudes entrantes para otorgar las reservas asociadas puede considerarse como un proceso de asignación de recursos a agentes que los solicitan, se supone que los gestores de intersección deben asignar los recursos disputados a los agentes que más los valoran. En línea con los hallazgos del diseño de mecanismos, se asume que cuanto más esté dispuesto a pagar un conductor por el conjunto deseado de espacios e intervalos de tiempo, más valora el bien. En este sentido, una política de asignación de recursos apropiada consiste en aplicar subastas combinatorias (*combinatorial auction*, CA). Los bienes asignados mediante estas subastas combinatorias son el derecho a usar cierto espacio dentro de la intersección en un momento dado. Una intersección se modela como una matriz discreta de espacios. Sea S el conjunto de espacios de la intersección y T el conjunto de pasos temporales futuros; entonces, el conjunto de ítems por los que un postor puede pujar es $I = S \times T$. Debido a la naturaleza del problema, un postor solo está interesado en paquetes de ítems sobre el conjunto I que reflejen su trayectoria deseada. Como ilustra la figura 1, en ausencia de aceleración en la intersección, una solicitud de reserva define implícitamente qué espacios y en qué momentos necesita el conductor para atravesar la intersección. Una puja sobre un paquete de ítems se define implícitamente por la solicitud de reserva. Dados los parámetros de hora de llegada, velocidad de llegada, carril y tipo de giro, el subastador (es decir, el gestor de la intersección) puede determinar qué espacios se necesitan en qué momento. Además, el valor de la puja, es decir, la cantidad de dinero que el conductor está dispuesto a pagar por la reserva solicitada, debe incluirse en la solicitud de reserva del vehículo.

La figura 2 muestra el protocolo de interacción utilizado para regular la subasta combinatoria. Comienza con el subastador esperando pujas durante un cierto tiempo. Una vez recopiladas las nuevas pujas (*bid set*), el subastador ejecuta un algoritmo de aproximación *anytime* para resolver el problema de determinación de ganadores

FIGURA 1

CONJUNTO DE ESPACIOS INCLUIDOS EN UNA SOLICITUD DE RESERVA

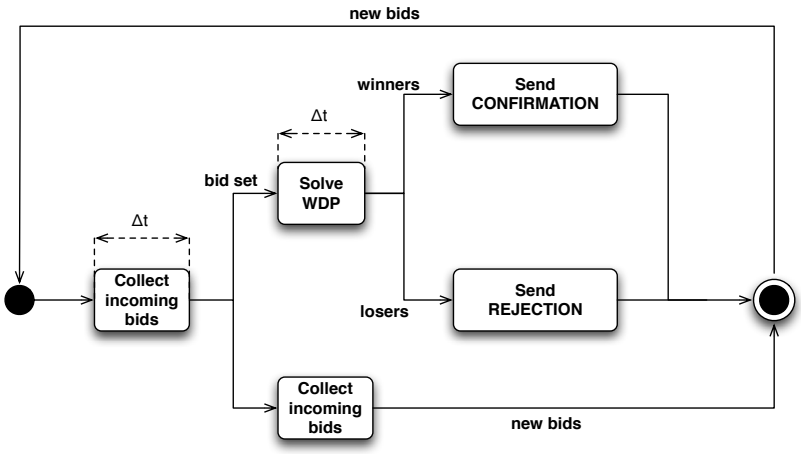


Fuente: Vasirani and Ossowski (2012).

(*winner determination problem*, WDP), determinando así el conjunto ganador con las pujas cuyas solicitudes de reserva han sido aceptadas. Durante la ejecución del algoritmo WDP, el subastador sigue aceptando pujas entrantes, pero estas solo se incluirán en el conjunto de pujas de la siguiente ronda. El subastador envía un mensaje de *CONFIRMATION* a todos los postores cuyas pujas están en el conjunto ganador, mien-

FIGURA 2

PROTOCOLO DE SUBASTA PROPUESTO



Fuente: Vasirani and Ossowski (2012).

tras que se envía un mensaje de *REJECTION* a los postores que enviaron las pujas restantes. Después comienza una nueva ronda y el subastador recopila nuevas pujas entrantes durante un cierto tiempo.

Mediante un conjunto de experimentos simulados, Vasirani y Ossowski (2012) confirmaron que la política basada en subastas combinatorias (CA) impone una relación inversa entre las cantidades gastadas por los postores y su retraso (el aumento en el tiempo de viaje debido a la presencia de la intersección). Es decir, cuanto más dinero esté dispuesto a gastar un conductor para cruzar la intersección, más rápido será su tránsito por ella. Notaron que, incluso con una cantidad teóricamente infinita de dinero, un conductor no puede experimentar un retraso cero al aproximarse a una intersección, ya que el tiempo de viaje está influenciado por vehículos más lentos y potencialmente más pobres que se encuentran delante.También confirmaron que la política CA tiene un retraso promedio mayor en comparación con la estrategia FCFS, ya que concede una reserva al conductor que más la valora, en lugar de maximizar el número de solicitudes concedidas. Este *coste social* de CA es mayor cuanto más alta es la carga de la intersección, es decir, cuantos más conductores compiten por las reservas. La extensión del mecanismo CA a múltiples intersecciones descrita a continuación tiene como objetivo reducir este coste social de dar preferencia a conductores con una alta valoración del tiempo.

2.3.3. Redes de intersecciones basadas en subastas

En el caso de una única intersección, un agente conductor simplemente necesita decidir el valor preferido de la puja que enviará al subastador. El espacio de decisión de un agente conductor en una red urbana con múltiples intersecciones es mucho más amplio: deben tomarse decisiones complejas y mutuamente dependientes, como la elección de ruta y la selección del momento de salida. Por lo tanto, este escenario abre nuevas posibilidades para que los gestores de intersección influyan en el comportamiento de los conductores. Por ejemplo, un gestor de intersección puede estar interesado en influir en la elección colectiva de rutas realizada por los conductores, utilizando paneles de mensajes variables, difusión de información o sistemas individuales de guía de rutas, con el fin de distribuir el tráfico de manera uniforme en la red. Este problema se denomina *asignación de tráfico*. En la estrategia "asignación competitiva de tráfico con subasta combinatoria" (*Competitive Traffic Assignment–Combinatorial Auction*, CTA-CA), cada gestor de intersección ejecuta las subastas combinatorias descritas en la sección anterior, pero incluye un precio base de reserva, es decir, anuncia un precio mínimo que cobra por las reservas que asigna. En cada instante t, puede haber diferentes precios de reserva $p^t(l)$ para cada uno de los enlaces entrantes l de la intersección. Los precios de reserva de las intersecciones están disponibles en tiempo real para los agentes conductores, quienes eligen sus rutas según sus preferencias personales sobre tiempos de viaje y los costes monetarios actuales. Cada gestor de intersección

compite con los demás por el suministro de las reservas que se comercializan. Aplica una regla simple de actualización del precio de reserva que busca atraer conductores a la intersección cuando la demanda de tráfico es baja y disuadirlos de elegir rutas que la atraviesen cuando está sobresaturada. El nuevo precio de reserva en el instante $t+1$ en el enlace l ($p^{t+1}(l)$) se calcula en base al precio actual ($p^{t}(l)$) y teniendo en cuenta la actual saturación de la intersección en el enlace l :

$$p^{t+1}(l) \leftarrow p^{t}(l) + p^{t}(l) \cdot \frac{z^{t}(l)}{s(l)} \qquad [1]$$

En esta fórmula, la constante $s(l)$ representa la cantidad óptima de vehículos que pueden cruzar la intersección desde el enlace l. La demanda excedente $z^{t}(l)$ expresa la diferencia entre la demanda total en el instante t y la oferta $s(l)$ en el enlace l (nótese que $z^{t}(l)$ se vuelve negativa cuando la oferta en el enlace l supera la demanda). Ajustando dinámicamente los precios de reserva de los gestores de intersección de la red, los flujos de tráfico se distribuyen mejor, evitando la sobresaturación de intersecciones críticas. Vasirani y Ossowski (2012) evaluaron el enfoque mediante simulaciones sobre una topología que se asemejaba a la red vial de alta capacidad de la ciudad de Madrid. Entre otros aspectos, compararon CTA-CA con una estrategia FCFS, donde cada gestor de intersección asigna reservas de forma aislada siguiendo el criterio de "primero en llegar, primero en ser atendido". Nótese que, a diferencia de CTA-CA, en FCFS no existe noción de coste social, ya que todos los conductores son tratados por igual; pero, por otro lado, FCFS no puede beneficiarse de los efectos de asignación de tráfico derivados de los precios de reserva de CTA-CA. En los experimentos, con la estrategia FCFS, el modelo de elección de ruta de los conductores simplemente selecciona la ruta con el tiempo de viaje esperado mínimo en condiciones de flujo libre, ya que no existe noción de precio. Para la evaluación de la estrategia CTA-CA, se asume que los conductores eligen la ruta más preferida que pueden permitirse. Los experimentos miden la media móvil del tiempo de viaje, es decir, cómo evoluciona el tiempo de viaje promedio de toda la población de conductores, calculado sobre todos los pares origen-destino, durante la simulación. Los resultados muestran que CTA-CA supera a FCFS, manteniendo la relación inversa entre las cantidades gastadas por los agentes conductores y el retraso de sus vehículos. La ganancia de rendimiento de CTA-CA es mayor cuanto más alta es la carga en las intersecciones. Estos resultados indican que los métodos de asignación preferente, como CTA-CA, pueden proporcionar un acceso individualizado a la infraestructura pública de tráfico adaptado a las preferencias y necesidades del conductor, mientras que su efecto negativo sobre el bienestar social (es decir, el aumento del tiempo de viaje promedio) puede mitigarse mediante un diseño adecuado de los mecanismos de asignación.

2.3.4. Cuestiones técnicas, legales y éticas

El enfoque arriba expuesto a modo de ejemplo se ha diseñado como prototipo para un hipotético futuro con coches autónomos guiados por agentes IA. Hay varias cuestiones técnicas abiertas y oportunidades de mejora. Por ejemplo, la eficiencia del mecanismo baja en la medida en la que crece la incertidumbre sobre el tiempo de llegada de los coches autónomos a las intersecciones. Por otra parte, sería interesante estudiar las posibilidades de las comunicaciones entre vehículos (V2V) para enriquecer el espacio de acción de un conductor, por ejemplo, mediante la opción de unirse o abandonar dinámicamente coaliciones de vehículos, basándose en la idea de pelotones de vehículos.

Sin embargo, aparte de estas cuestiones meramente técnicas, hay una serie de retos legales y éticos que habría que atacar para que, en este futuro hipotético ya mencionado, y una vez superadas cuestiones técnicas abiertas, se pudiera plantear la implantación de un mecanismo de estas características. Para identificar estas cuestiones, se ha aplicado el método descrito en (Santos *et al.*, 2020) y resumido en el apartado 2.2. Como conclusiones de este análisis, se identifican varios puntos que requieren modificaciones técnicas o legales (en relación a la legislación española).

Como punto más obvio, la propuesta planteada se basa en el uso de vehículos autónomos que actualmente no están permitidos en España, por lo que desde un punto de vista legal se requeriría un cambio normativo en este aspecto, mientras que desde una perspectiva técnica haría falta una extensión del mecanismo que, al menos durante un tiempo, permita la coexistencia de vehículos dirigidos por agentes IA y por conductores humanos. De igual forma, será necesario cambiar la legislación de tráfico para contemplar los gestores de intersección.

Por otra parte, el hecho de que un vehículo deba solicitar franjas de espacio y tiempo en un momento determinado al gestor de la intersección correspondiente no es contrario a la ley. Ni tampoco lo es que estas franjas se asignen en base a una subasta electrónica de puja única y cerrada. Desde la perspectiva legal, las subastas electrónicas ya están reguladas[3], por lo que en este aspecto el mecanismo propuesto cumpliría con la legislación vigente.

Otra cuestión analizada plantea que, en el sistema propuesto, el gestor de intersección puede obtener beneficios. Desde el punto de vista legal, la implementación se puede establecer de dos maneras: mediante gestión pública sin beneficios, o mediante gestión privada con beneficios a través de una concesión. Ambas alternativas cumplirían con la legislación vigente.

[3] Ley 33/2003, de 3 de noviembre, del Patrimonio de las Administraciones Públicas, y Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público.

Asimismo, los gestores de intersección establecen precios de reserva principalmente para mejorar la distribución del tráfico, más que con el objetivo de obtener beneficios. No obstante, el precio de la reserva no está limitado y fluctúa libremente con la demanda de tráfico. Según la legislación vigente, la Administración Pública es competente para determinar un precio máximo. Por tanto, se requeriría una modificación técnica que tenga en cuenta estos precios máximos.

Finalmente, para participar en la subasta, cada intersección tiene un precio de reserva que ha sido publicado por el gestor de la intersección y la oferta de cada vehículo debe ser superior al precio de reserva actual de la intersección. Aquí, la propuesta no cumple con la legislación, ya que las carreteras son bienes de dominio público y los conductores deben tener la opción de utilizarlas de forma gratuita. En este sentido, se requiere un cambio técnico que asegure que un conductor pueda obtener un acceso gratuito a la intersección. Este cambio es técnicamente factible sin perder la eficiencia del sistema, por ejemplo, al permitir el pase gratuito por una intersección a conductores que hayan esperado un tiempo determinado.

Aparte de estas cuestiones de legalidad, hay una serie de aspectos éticos y morales que deben ser tenidos en cuenta para que la propuesta sea aceptable. El hecho de "mercantilizar" el derecho a los servicios básicos de una infraestructura pública, mencionada en el párrafo anterior, es solo uno de ellos. También se debe determinar el comportamiento del sistema y las responsabilidades si se producen fallos en el mismo y asegurar la protección de datos privados. Y finalmente, se debe tener en cuenta la posibilidad de que ciertos vehículos deben tener prioridad a la hora de pasar por una intersección (por ejemplo, ambulancias, coches de policía o bomberos), o podrían tener ciertas preferencias (transporte público o vehículos de personas discapacitadas, por ejemplo).

3. AGENTES Y VALORES HUMANOS

En la medida en que aumenta el nivel de abstracción en la definición de las tareas que se delegan en los agentes IA dentro de un sistema sociotécnico, también crecen de forma sustancial el número y la complejidad de las decisiones que estos han de tomar de forma autónoma. Como se ha argumentado en la sección anterior, es necesario gobernar este tipo de sistemas tanto para asegurar una adecuada coordinación en la interacción entre los agentes, como para garantizar su conformidad con la legislación. Sin embargo, para fomentar la aceptación de este tipo de sistemas en la sociedad, esto no es suficiente. Podría entenderse como un caso aislado, pero, de hecho, al disponer de mayor autonomía, los aspectos éticos de las decisiones que toman los agentes IA son más relevantes de lo que pueden parecer. Una de las razones es que un agente IA se puede encontrar ante *dilemas morales*.

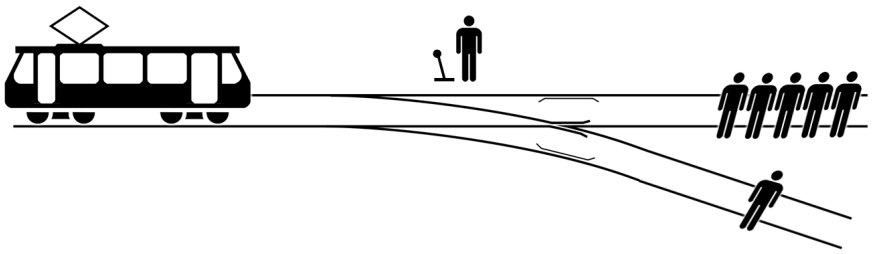
A partir de este concepto, en el apartado 3.1 se motiva la necesidad de desarrollar agentes IA capaces de razonar sobre los aspectos éticos de sus decisiones. El apartado 3.2 presenta un modelo para representar formalmente sistemas de valores éticos. Finalmente, en el apartado 3.3 se plantea un método para aprender sistemas de valores a partir de ejemplos de las preferencias que los agentes de una sociedad mantienen en relación con las diferentes alternativas de decisión en un dominio concreto.

3.1. Agentes basados en valores

Los dilemas morales son ampliamente estudiados en Ética, Derecho, Psicología y otras disciplinas. Esencialmente, se refieren a situaciones en las que hay dos o más imperativos morales. El ejemplo clásico de una situación de estas características es el llamado dilema del tranvía, inicialmente ideado por la filósofa Philippa Foot, y más tarde planteado en su forma actual por Judith Jarvis Thomson (Thomson, 1976). La figura 3 ilustra el relato[4]: un tranvía corre fuera de control por una vía. En su camino se hallan cinco personas que acabarían atropelladas. Es posible accionar una palanca que encaminará al tranvía por una vía diferente, pero, por desgracia, hay otra persona en esta que también moriría atropellada. La cuestión es, si pudiéramos accionar la palanca, ¿lo deberíamos hacer? El dilema moral que plantea esta pregunta es decidir entre causar un mal o dejar que ocurra otro, potencialmente mayor. Desde un punto de vista utilitario, deberíamos accionar la palanca, ya que de esta forma salvaríamos la vida de cinco personas. Pero desde otro punto de vista, es moralmente deplorable tomar acciones que dañen a una persona, lo cual aconsejaría la inacción y dejar que el tranvía siga recto.

FIGURA 3

DILEMA DEL TRANVÍA



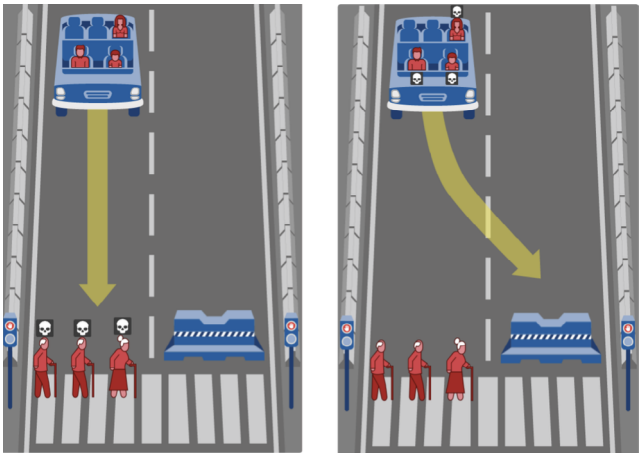
Fuente: Wikipedia.

[4] https://es.wikipedia.org/wiki/Dilema_del_tranvía

A primera vista, el ejemplo parece artificial, pero no es descartable que un agente IA, debido a su autonomía, también se encuentre alguna vez en una situación que requiera decisiones de estas características. Esto se ilustra muy bien en el ámbito de la conducción autónoma a través del llamado *experimento de la Máquina Moral*[5], publicado en un artículo en la revista *Nature* (Awad *et al.*, 2018). En este artículo, los autores presentaban a sujetos de diferentes países y culturas situaciones en las que un vehículo autónomo tenía que elegir entre dos daños inevitables. La figura 4 muestra un ejemplo: un coche conducido por un agente IA se acerca a un paso por el que transitan varias personas, y sufre un fallo repentino en sus frenos. Al agente solo le quedan dos opciones: o bien seguir adelante, en cuyo caso atropellaría a los transeúntes, o bien hacer una maniobra brusca de evasión a consecuencia de la cual el coche chocaría contra un obstáculo, lo que causaría la muerte a todos sus pasajeros. Los sujetos tenían que articular qué elección pensaban que debía tomar el agente IA del coche autónomo. Los autores registraban estas decisiones y las analizaban posteriormente.

FIGURA 4

ESCENARIO DEL EXPERIMENTO DE LA MÁQUINA MORAL



Fuente: Awad *et al.* (2018).

Las elecciones de las personas en situaciones con alta carga ética, como los dilemas morales, se basan en sus *valores morales*. Estos valores les permiten discernir entre lo que consideran el bien y el mal, y decidir qué tipo de situaciones se deberían promocionar. Los humanos a menudo mantenemos múltiples valores morales, que muchas veces están compitiendo entre sí como, por ejemplo, el valor de la meritocracia y el valor de la igualdad. Un *sistema de valores* determina la relativa importancia de cada uno de los

[5] https://www.moralmachine.net/hl/es

valores. Por ejemplo, una persona puede considerar que, según su sistema de valores, la meritocracia es más relevante que la igualdad.

Muchos valores están compartidos entre diferentes sociedades y culturas, pero sus sistemas de valores, es decir, la importancia relativa que se da a cada valor, suelen ser diferentes. A partir de los datos obtenidos con diferentes escenarios en el experimento de la Máquina Moral, se identificaron tres grupos principales de sujetos que compartían sistemas de valores similares. A grandes rasgos, ante un dilema moral para el agente IA de un coche autónomo como el indicado anteriormente, los sujetos pertenecientes a países occidentales en general preferían la inacción. Para los sujetos procedentes de países orientales era importante si tanto el coche autónomo como los transeúntes se comportaban de acuerdo con la ley o no. Finalmente, los sujetos de países sureños juzgaban relevantes también otras características como el estatus social o el género de los individuos afectados.

Como sociedad, deseamos asegurar que las acciones de los agentes IA, y las consecuencias de estas, estén acordes a nuestro sistema de valores (conocido como el *problema de alineamiento con los valores*). Ciertamente, se puede cuestionar si un agente IA que conduce un coche autónomo ha de guiarse directamente por el sistema de valores de la sociedad en la que está embebido, es decir, el sistema sociotécnico correspondiente. Pero parece evidente que, si ignora completamente este sistema de valores, su nivel de aceptación será limitado.

Desde el punto de vista de un ingeniero, por tanto, es relevante tener herramientas que, una vez identificado el sistema de valores por el que un agente IA se debe guiar, aseguren que sus decisiones y acciones estén alineadas con él. Una vía para afrontar este problema es durante el proceso de diseño a través de una metodología como el diseño sensitivo a los valores (*Value Sensitive Design*, VSD) (Friedman *et al.*, 2013). Una alternativa más flexible es desarrollar representaciones formales de sistemas de valores, que permitan a un agente IA razonar con y sobre ellos. Estos "agentes conscientes de los valores" (*value-aware agents*) (Osman, 2025) serían capaces de asegurar un adecuado nivel de alineamiento de sus acciones en tiempo de ejecución razonando, si fuese necesario, sobre la adecuación ética de las diferentes decisiones que se les planteen en cada momento, y explicándolas en términos entendibles por los humanos. Se va a seguir este último enfoque en el resto de esta sección.

3.2. Modelado de sistemas de valores

Para desarrollar un modelo del sistema de valores de una sociedad, primero hay que identificar los valores relevantes para sus miembros. Varias teorías abogan por la existencia de un conjunto de valores universales, aplicables en todos los contextos y asumidos en todas las culturas. Entre las propuestas más conocidas se encuentra la teoría de

los valores universales de Schwartz (1992) que postula la existencia de diez tipos básicos de valores motivacionales presentes en todas las culturas (Benevolencia, Universalismo, Autodirección, Estimulación, Hedonismo, Logro, Poder, Seguridad, Conformidad y Tradición). Schwartz los organiza en un círculo según sus relaciones de compatibilidad y conflicto. Estos valores son creencias sobre estados deseables que trascienden situaciones específicas y guían la conducta de los individuos en una sociedad. Otro ejemplo es la teoría de los fundamentos morales (*Moral Foundations Theory*, MFT) procedente de la psicología social que propone un marco para explicar cómo la moralidad humana surge de "fundamentos" innatos y universales (Haidt and Joseph, 2004; Graham *et al.*, 2013). Otros autores abogan por la existencia de valores específicos que son relevantes en determinados contextos, y que van emergiendo continuamente en diferentes dominios (Van de Poel, 2021). De forma similar, la metodología del diseño sensitivo a valores, mencionada anteriormente, por ejemplo, argumenta que los valores son conceptos abiertos que deben ser extraídos por las partes interesadas en cada dominio específico.

Holgado-Sánchez *et al.* (2025, 2026) definen un modelo formal en el que ambos enfoques pueden ser retratados partiendo de un conjunto finito $V = \{v_1, \ldots, v_m\}$ de "etiquetas" v_i que representan *valores abstractos* y que toman un significado específico en un contexto. Por ejemplo, en el contexto del transporte, los valores relevantes podrían ser *eficiencia* (Eff), *sostenibilidad* (Sos) y *seguridad* (Seg). Comúnmente, un *sistema de valores* (abstracto) se modela a través de algún tipo de relación de preferencia sobre V . Por ejemplo, el sistema de valores de un agente puede indicar que para él la *sostenibilidad* es más importante que la *seguridad*, y esta a su vez es más importante que la *eficiencia*, es decir $Sos \succcurlyeq Seg \succcurlyeq Eff$.

Para que este sistema de valores abstracto adquiera un significado práctico, hay que interpretarlo en un dominio concreto. En particular, si en un dominio un agente ha de elegir entre diferentes alternativas, el sistema de valores induce un orden débil de preferencia sobre éstas alternativas. Por ejemplo, si para realizar un viaje se puede elegir entre $Metro$, $Coche$ o $Bici$, el sistema de valores abstracto VS arriba indicado podría implicar que $Metro \succcurlyeq_{VS} Bici \succcurlyeq_{VS} Coche$. Para poder razonar de forma automática en base a un sistema de valores abstractos, el modelo ha de especificar cómo se deduce esta relación de preferencia sobre las alternativas del dominio a partir de este sistema.

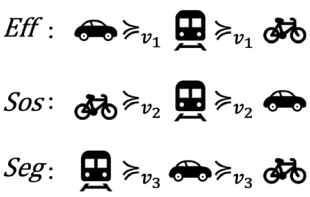
Para ello, primero hay que dar un significado concreto a cada valor en el dominio. Se considera que cada valor v_i establece un orden débil de preferencias $\preccurlyeq_{v_i}$ sobre las alternativas del dominio. La figura 5a muestra un ejemplo para la elección del medio de transporte: con respecto a la eficiencia, se prefiere el coche sobre el metro sobre la bici, con respecto a la sostenibilidad la bici sobre el metro sobre el coche, y con respecto a la seguridad el metro sobre el coche sobre la bici. Una *función de alineamiento con un*

valor A_v asigna un número a cada una de las opciones, de tal forma que este número es más alto cuanto más preferida sea una opción con respecto al valor v. Lo importante es que esta función ha de ser computacional, es decir, se debe poder calcular a partir de características de cada alternativa. Por ejemplo, la función de alineamiento con el valor de sostenibilidad se podría definir en base a la cantidad de CO_2 que produce cada una de las alternativas, de modo que sea inversamente proporcional a esta característica. El conjunto G_V de todas las funciones de alineamiento con un valor se llama *grounding* del sistema de valores en un dominio, ya que da un significado concreto y computable a todas las (etiquetas de) valores del sistema de valores abstracto. La figura 5b muestra un *grounding* compatible con el ejemplo. Muchas veces se asume que este *grounding* de los valores es compartido entre todos los agentes de una sociedad, es decir, que todos los individuos entienden (aproximadamente) lo mismo con respecto a lo que significa que un medio de transporte sea más o menos eficiente, seguro y sostenible.

FIGURA 5

VALORES Y *GROUNDING* G_v

A)



B)

A_v	*Eff*	*Sos*	*Seg*
	10	2	9
	7	8	10
	4	10	6

Fuente: Elaboración propia.

El significado del sistema de valores de un agente j en un dominio se define de forma similar ya que implica un orden débil de preferencias $\preccurlyeq^j_{G_V}$ sobre las alternativas, en este caso reflejando las preferencias particulares del agente j. La *función de alineamiento con el sistemas de valores* A_{f_j,G_V} de un agente j se calcula a partir del *grounding* G_V con una función de agregación f_j, de tal forma que representa el orden débil correspondiente. Es decir, para todas las alternativas e y e' se cumple

$$A_{f_j,G_V}(e) \leq A_{f_j,G_V}(e') \Leftarrow e \preccurlyeq^j_{G_V} e' \qquad [2]$$

A menudo se modela f_j como una función lineal, dando diferentes pesos w_i a cada valor v_i en una suma ponderada. Los pesos se pueden interpretar como la importancia que un agente j da a cada uno de los valores en su sistema de valores. Por ejemplo, el agente j podría dar un peso de 0.2 al valor de eficiencia, un peso de 0.5 al valor de

sostenibilidad, y un peso de 0.3 al valor de seguridad. En consecuencia, sobre la base del *grounding* de la figura 5b, obtendríamos $A_{f_j,G_V}(Coche) = 5.7$; $A_{f_j,G_V}(Metro) = 8.4$; y $A_{f_j,G_V}(Bici) = 7.6$; lo cual coincide con el orden de preferencias indicado al inicio del ejemplo. Nótese que, a pesar de que en el sistema de valores abstracto el valor más preferido es la *sostenibilidad* (Sos), en el dominio del ejemplo la opción más preferida es $Metro$, aunque es menos preferida que $Bici$ con respecto al valor Sos (ver figura 4a). Sin embargo, respecto a los dos valores Eff y Seg la opción $Metro$ se prefiere sobre $Bici$ lo cual, dados los pesos que definen la función de agregación, compensa la preferencia de $Bici$ sobre $Metro$ respecto al valor Sos .

Se ha definido lo que es un sistema de valores de un agente, pero falta formalizar qué se entiende por un sistema de valores de una sociedad. La solución más inmediata es concebirlo simplemente como una agregación (un promedio) de los sistemas de valores de los miembros de la sociedad. Sin embargo, esto podría no reflejar adecuadamente la diversidad existente en ella. Por ejemplo, una sociedad con dos grupos claramente diferenciados y de tamaño similar (p.ej., "izquierdistas" y "derechistas") es claramente distinta de una sociedad donde todos los miembros comparten un sistema de valores intermedio ("centristas"). Para tener en cuenta esta peculiaridad, en vez de representar las preferencias éticas de una sociedad a través de un único sistema de valores, Holgado-Sánchez *et al.* proponen modelarlas mediante un conjunto de sistemas de valores: uno para cada subgrupo "relevante" que exista en la sociedad (Holgado-Sánchez *et al.*, 2025).

3.3. Aprendizaje de sistemas de valores

Uno de los mayores obstáculos para la aplicación práctica de un modelo computacional de sistemas de valores, como el que se ha esbozado en la sección anterior, reside en la dificultad de identificar tanto la interpretación o el significado de un conjunto de valores como los sistemas de valores concretos que existan en una sociedad, en un dominio de interés concreto. Se puede obtener esta información y codificarla "a mano" (por ejemplo, a partir de encuestas o entrevistas con las personas involucradas), pero este enfoque es delicado por varias razones: i) los valores son conceptos abstractos cuya interpretación varía entre distintos contextos y grupos sociales; ii) las personas generalmente no tienen una idea clara de qué valores realmente influyen en sus decisiones (no suelen tener una idea clara de su propio sistema de valores); y iii) la dificultad en la calibración de estas instancias de sistemas de valores.

Para solventar este problema, en (Liscio *et al.*, 2023) se propone un proceso genérico denominado *inferencia de valores* para identificar el sistema de valores de una sociedad de agentes. El proceso se estructura en tres fases:

1. Identificación de valores, es decir, el proceso de encontrar el conjunto de valores $V = \{v_1, \ldots, v_m\}$ relevante para un contexto determinado. Este proceso requiere llevar a cabo conversaciones con expertos del dominio, encuestas o un análisis de documentos mediante técnicas de procesado de lenguaje natural.

2. Estimación de sistemas de valores, es decir, el proceso de identificar los sistemas de valores de los agentes de la sociedad. El problema típicamente se traduce en la especificación de un orden débil de preferencia entre los valores, o bien un sistema de pesos que represente la importancia de cada valor para cada individuo del sistema modelado, como se ha visto en la sección anterior.

3. Agregación de sistemas de valores, es decir, el problema de encontrar una representación consensuada del sistema de valores de la sociedad, por ejemplo, mediante técnicas de elección social basadas en consenso.

Un aspecto clave en el proceso de inferencia de valores es la especificación del significado de cada valor en términos de las características del dominio. Un posible enfoque consiste en concebir esta tarea como un problema de *aprendizaje de valores* (Soares, 2019). Holgado-Sánchez *et al.* siguen este enfoque y, asumiendo el modelo descrito en el apartado 3.2, proponen aprender tanto el significado de un conjunto de valores abstractos como los sistemas de valores de agentes (Holgado-Sánchez *et al.*, 2025). En otro trabajo (Holgado-Sánchez *et al.*, 2026), se extiende este planteamiento al aprendizaje del sistema de valores de una sociedad.

En estos artículos, una vez identificado un conjunto de valores $V = \{v_1, \ldots, v_m\}$ mediante técnicas de *identificación de valores*, el problema de aprendizaje de valores consiste en aprender un *grounding* G_V que aproxime la conceptualización de los valores por parte de los agentes de la sociedad. Para ello, se plantea aprender las funciones de alineamiento A_{v_i} a partir de evaluaciones (no necesariamente cuantificadas) entre alternativas basadas en cada uno de los valores. Es decir, los conjuntos de datos de entrada consisten en pares de alternativas, junto con la valoración de un agente sobre cuál de estas alternativas está más alineada con cada valor (según su entendimiento). Estos datos se obtienen a través de observaciones y/o mediante interacción directa con los agentes de la sociedad. Así se evita tener que recurrir a procedimientos más costosos como encuestas o estimaciones directas de funciones de alineamiento.

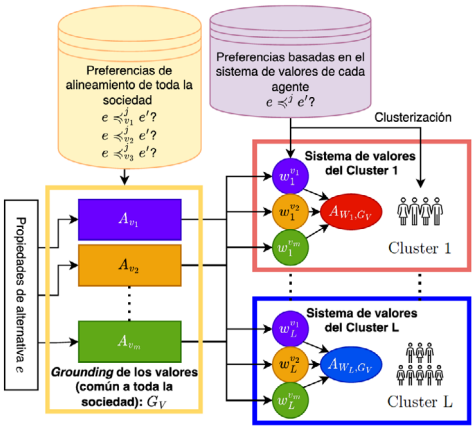
Además, siguiendo un proceso similar, en (Holgado-Sánchez *et al.*, 2025) se propone aprender una representación de los sistemas de valores de la sociedad: dados ejemplos de las preferencias entre alternativas según el propio sistema de valores de cada agente (es decir, tomando sus preferencias tras ponderar todos los valores), se puede estimar una representación de la importancia general que cada valor tiene en esa sociedad para distintos grupos de agentes.

Mediante técnicas de regresión logística (profunda), se estima un modelo para cada uno de estos conjuntos de datos y agentes. El problema principal consiste en cómo agrupar los agentes en distintos *clústers* de forma que todos ellos vean representado "suficientemente bien" su sistema de valores mediante un sistema de valores común (un mismo sistema de pesos definido sobre el *grounding* compartido). Un conjunto de *clústers* con sus respectivos sistemas de valores se define como el *sistema de valores de una sociedad*.

La figura 6 resume el esquema de aprendizaje. Cada alternativa se evalúa mediante una función de alineamiento con cada valor (A_{v_i} , $i = 1,...,m$), que constituyen un *grounding* de dichos valores. Instancias de este *grounding* se aprenden a partir de ejemplos de valoraciones entre alternativas, en los que los agentes j de la sociedad indican qué alternativa está más alineada con cada valor. Por otro lado, el sistema de valores de la sociedad se define como un conjunto de L sistemas de valores definidos, y su asignación a distintos grupos de agentes (*clústers*). Siguiendo el modelo descrito en el apartado 3.2, estos sistemas de valores se definen a su vez mediante distintos conjuntos de pesos normalizados ($W_l = (w_l^{v_1},\ldots,w_l^{v_m})$ con $w_l^{v_i} \in [0,1]$ y $\sum_{i=1}^{m} w_l^{v_i} = 1$) que estiman la importancia relativa de cada valor. La función de alineamiento del sistema de valores de cada *clúster* evalúa cada alternativa con una agregación lineal del *grounding* ponderada con estos pesos. Es decir, para cada *cluster* l, su función de alineamiento está dado por $A_{W_l,G_V}(e) = W_l \cdot G_V(e)^T$. Tanto el reconocimiento de *clústers* como el aprendizaje de los conjuntos de pesos correspondientes a cada uno de los mismos se realizan a partir de la observación de ejemplos de preferencias entre alternativas, en los que cada agente de la sociedad expresa qué alternativa está más alineada sobre la base de su propio sistema de valores.

FIGURA 6

REPRESENTACIÓN DEL MODELO SISTEMA DE VALORES DE UNA SOCIEDAD Y DE LOS DATOS NECESARIOS PARA EL APRENDIZAJE DE UNA INSTANCIA DEL MISMO



Fuente: Elaboración propia.

Durante el proceso de aprendizaje se buscan funciones de alineamiento A_{v_i} , *clústers* de agentes, y conjuntos de pesos W_l , tal que se maximizan tres métricas cuantitativas:

1. Coherencia del *grounding*: expresa en qué medida se respeta el entendimiento de los valores de cada agente, es decir, en qué grado las funciones de alineamiento representan los ejemplos de preferencias de cada agente.

2. Representatividad: determina en qué medida es representado el sistema de valores individual de cada agente por el sistema de valores de su *clúster*. Se mide en función del número de coincidencias entre las preferencias indicadas de cada agente y la función de alineamiento del sistema de valores del *clúster* al que es asignado.

3. Concisión: mide la complejidad de la representación de sistemas de valores obtenida. Formalmente, la concisión se determina mediante las distancias entre las preferencias que inducen los sistemas de valores aprendidos en el espacio de alternativas, y se quiere maximizar para garantizar que la existencia de cada uno de los sistemas de valores sea realmente relevante. Maximizando la concisión, también se tiende a minimizar el número de *clústers*.

En Holgado-Sánchez *et al.* (2025), se evalúa el esquema de aprendizaje propuesto usando un conjunto de datos reales sobre la elección de rutas de tren en Suiza, en el que participaron 388 personas (agentes) (Vrtic and Axhausen, 2002). En este conjunto de datos, cada agente indica su ruta preferida entre dos alternativas en nueve situaciones distintas (es decir, un total de 3.492 observaciones). Cada ruta se describe mediante cuatro atributos: tiempo de viaje, coste, número de transbordos necesarios, y tiempo de espera entre trenes. Además, en este conjunto de datos se recogieron seis características contextuales de cada agente: nivel de ingresos del hogar, disponibilidad de coche, y el propósito del viaje (desplazamiento al trabajo, compras, negocios u ocio, siendo estos últimos excluyentes entre sí). En el análisis, cada ruta se considera una opción posible y el conjunto de participantes forma la población estudiada. Los datos reflejan únicamente qué ruta prefiere cada agente en cada comparación, es decir, qué ruta de cada par escogería el agente.

Se asume que las decisiones sobre las rutas están guiadas por tres valores principales: *eficiencia temporal*, *eficiencia económica*, y *comodidad*. La eficiencia temporal y la económica se definen directamente a partir del tiempo de viaje y del coste, respectivamente. En cambio, la comodidad se relaciona con el tiempo de espera entre trenes y el número de transbordos: una ruta se considera más cómoda si tiene menor tiempo de espera y menos transbordos. Cuando solo uno de estos dos aspectos mejora y el otro no, no se establece una preferencia entre las rutas y se deja que el modelo aprenda libremente cómo se valora la comodidad. A partir de estas definiciones, se construye el conjunto de datos de valoraciones de alternativas para el aprendizaje de las funciones de alineamiento A_{v_i} a partir de los 3.492 pares de alternativas originales.

Al utilizar los algoritmos de aprendizaje sobre este conjunto de datos, se obtuvo una solución con tres sistemas de valores para sendos grupos de agentes (*clústers*). Estos *clústers* exhibieron ciertas regularidades interesantes, como por ejemplo que los agentes que fueron asignados a un sistema de valores que daba la mayor importancia a la eficiencia económica, solían elegir los trayectos con la intención de realizar compras en el destino, mientras que los agentes que preferían el valor de la eficiencia temporal típicamente realizaban trayectos con fines de negocios. Esa información contextual (viajes de compras/negocios) no se tuvo en cuenta en el proceso de aprendizaje y *clusterización*, lo cual resalta que el modelo y algoritmos empleados no solo maximizan métricas cuantitativas sino también reconocen patrones implícitos de preferencias.

4. CONCLUSIONES

Una característica clave de los agentes IA es su capacidad para tomar decisiones y actuar de forma *autónoma* en los entornos en los que están embebidos. Generalmente, para alcanzar sus objetivos y realizar las tareas que les han sido encomendadas, los agentes IA han de interactuar con otros agentes y, a veces, también directamente con las personas. En este capítulo se ha argumentado que una *gobernanza* efectiva para los sistemas multiagente y sociotécnicos resultantes es fundamental para que las aplicaciones basadas en IA sean seguras, responsables y confiables.

Se ha razonado que, para alcanzar una coordinación adecuada entre los agentes IA, los diseñadores han de actuar sobre su entorno, y para muchos problemas esto se traduce en imponer unos protocolos de interacción adecuados sobre las infraestructuras de comunicación que facilitan el intercambio de mensajes de los agentes. Además, se ha ilustrado esta idea con un ejemplo de redes de gestores de intersección basados en reservas, para un hipotético futuro con coches autónomos guiados por agentes IA. También se ha comentado que, aparte de una serie de problemas técnicos, enfoques de estas características tienen implicaciones legales que se han de afrontar tanto desde un punto de vista técnico como regulatorio.

Por otra parte, se ha ilustrado que los problemas éticos en sistemas sociotécnicos con agentes IA son más relevantes de lo que pueda aparentar a primera vista. En este sentido, es deseable asegurar el alineamiento de las acciones de los agentes IA con el sistema de valores de la sociedad; un objetivo que se puede conseguir a través del concepto de los *agentes conscientes de los valores*. Para ello, se ha esbozado un enfoque computacional para modelar el significado de valores éticos en dominios concretos, y cómo se puede construir a partir de allí un modelo formal del sistema de valores de un agente o de una sociedad. Para instanciar este tipo de modelos, se ha presentado una propuesta para aprender el significado de los valores abstractos (su semántica computacional), así como el sistema de valores de una sociedad de agentes.

Quedan varias líneas de investigación abiertas en el ámbito del aprendizaje de sistemas de valores que merecen ser exploradas. Una de ellas es la ampliación de los métodos de aprendizaje para su aplicación en entornos más complejos con decisiones *secuenciales*. Partiendo del modelo arriba expuesto, en (Holgado-Sánchez *et al.*, 2026) se presenta y evalúa un método para ello, pero manteniendo la asunción de una función de agregación lineal. Otra línea consiste en hacer explícita la dependencia del *contexto* de los valores y de los sistemas de valores de los agentes, es decir, reconocer situaciones donde el sistema de valores de un agente cambie radicalmente ante cambios contextuales, como podrían ser casos de emergencia o durante interacciones con otros agentes. Por otro lado, en las propuestas presentadas se asume que exista un cierto consenso entre los individuos de la sociedad acerca del significado o de la interpretación de cada valor. Por ejemplo, eficiencia temporal, tiene una clara interpretación común: llegar antes al destino. En este sentido, puede ser de interés estudiar en más detalle qué dificultades o implicaciones puedan existir en sociedades más *heterogéneas*, es decir, en las que se consideren valores cuya interpretación varíe radicalmente entre grupos, como, por ejemplo, valores de *justicia* o *libertad*.

Otra línea de investigación relevante en este ámbito es el estudio de la relación entre valores y *políticas regulatorias*, donde estas últimas se caracterizan a través de los reglamentos y normas que las implementan. Para ello, conviene partir del gran abanico de enfoques propuestos en las últimas décadas en el ámbito de los sistemas multiagente para modelar normas, tanto sociales como legales. El trabajo de Montes y Sierra (2022) es pionero en este ámbito al proponer métodos para la síntesis automatizada de conjuntos de normas (políticas) que promocionan el alineamiento de un sistema multiagente con un conjunto de valores. Un dominio interesante para demostrar la relevancia social de este tipo de enfoques es el dominio de la asignación de plazas escolares. En España, una ley estatal regula el marco general para el proceso de asignación, pero, al ser un país políticamente descentralizado, varias comunidades autónomas han establecido sus propios reglamentos para la implementación efectiva de este proceso. Los resultados de dichos reglamentos son públicamente accesibles bajo la Ley de Transparencia (Moreno Rebato and Rodríguez García, 2025) y constituyen un cuerpo de datos valioso para analizar los sistemas de valores subyacentes a los diferentes reglamentos existentes, es decir, analizar qué tipo de valores se promueven en unos casos o en otros.

Referencias

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J-F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Dötterl, J., Bruns, R., Dunkel, J., and Ossowski, S. (2020). On-time delivery in crowdshipping systems: An agent-based approach using streaming data. In *Proc. Europ. Conf. on Artificial Intelligence (ECAI'20)*, 51–58. IOS Press.

Dresner, K., and Stone, P. (2008). A multiagent approach to autonomous intersection management. *J. Artif. Int. Res.*, *31*(1), 591–656.

Friedman, B., Kahn, P. H., Borning, A., and Huldtgren, A. (2013). *Value Sen-sitive Design and Information Systems*, 55–95. Dordrecht: Springer Netherlands.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine and A. Plant (editors), *Advances in Experimental Social Psychology*, volume 47, (55–130). Academic Press.

Haidt, J., and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4),55–66.

Holgado-Sánchez, A., Billhardt, H., Fernández, A., and Ossowski, S. (2026). Learning the value systems of agents with preference-based and inverse reinforcement learning. *Auton. Agent Multiagent Syst., 40*(4).

Holgado-Sánchez, A., Billhardt, H., Ossowski, S. and Degli-Esposti, S. (2025). Learning the value systems of societies from preferences. In *Proc. Europ. Conf. on Artificial Intelligence (ECAI'25)*, 1121–1130.

Holgado-Sánchez, A., Vamplew, P., Dazeley, R., Ossowski, S., and Billhardt, H. (2026). Learning the value systems of societies with preference-based multi-objective reinforcement learning. In *Proc. Int. Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS'26)*. To appear.

Liscio, E., Lera Leri, R. X., Bistaffa, F., Dobbe, R. I. J., Jonker, C. M., López-Sánchez, M., Rodríguez-Aguilar, J. A., and Murukannaiah, P. K. (2023). Value inference in sociotechnical systems. In *Proc. Int. Conf.on Autonomous Agents and Multiagent Systems (AAMAS'23)*, 1774–1780.

Montes, N., and Sierra, C. (2022). Synthesis and properties of optimally value-aligned normative systems. *J. Artif. Intell. Res.*, 74, 1739–1774.

Moreno Rebato, M., and Rodríguez García, J. A. (2025). El acceso a los datos en los procesos de asignación de plazas escolares con fines de investigación científica: La utilización de la inteligencia artificial en este contexto. *Revista Española de la Transparencia*, 21, 129–156.

Namazi, E., Li, J., and Lu, CH. (2019). Intelligent intersection management systems considering autonomous vehicles: A systematic literature review. *IEEE Access*, 7, 91946–91965.

Osman, N. (2025). Value-aware multiagent systems. In S. Cranefield, L. G. Nardin, and Lloyd N. (editors), *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of MultiAgent Systems XVII*, 32–37. Springer Nature Switzerland.

Sandholm, T. W. (1999). *Distributed rational decision making*, 201–258. Cambridge, MA, USA: MIT Press.

Santos, J-A., Fernández, A., Moreno-Rebato, M., Billhardt, H., Rodríguez-García, J. A., and Ossowski, S. (2020). Legal and ethical implications of applications based on agreement technologies: The case of auction-based road intersections. *Artif. Intell. Law, 28*(4), 385–414.

Schumacher, M., and Ossowski, S. (2006). The governing environment. In D. Weyns, H. Van Dyke Parunak, and F. Michel (editors), *Environments for Multi-Agent Systems II*, 88–104. Springer.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, 1–65. Elsevier.

Soares, N. (2019). The Value Learning Problem. In R. V. Yampolskiy (editor), *Artificial Intelligence Safety and Security*, chapter 7, 89–98. CRC Press.

Thomson, T. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217.

Van de Poel, I. (2021). Design for value change. *Ethics and Information Technology, 23*(1), 27–31.

Vasirani, M., and Ossowski, S. (2012). A market-inspired approach for intersection management in urban road traffic networks. *J. Artif. Int. Res., 43*(1), 621–659.

Vasirani, M., and Ossowski, S. (2013). A proportional share allocation mechanism for coordination of plug-in electric vehiclecharging. *Engineering Applications of Artificial Intelligence, 26*(3), 1185–1197.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance, 16*(1), 8–37.

Vrtic, M., and Axhausen, K. W. (2002). The impact of tilting trains in Switzerland. a route choice model of regional and long distance public transport trips. Technical report, IVT - ETH Zurich. DOI: 10.3929/ethz-a-004403563.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*, 2nd edition. Wiley.

CAPÍTULO VII

Máquinas que saben lo que importa: ingeniería de la conciencia de valores sociales

Carles Sierra*

La integración ética de la inteligencia artificial (IA) en la sociedad requiere superar tanto el catastrofismo apocalíptico como las limitaciones de los enfoques éticos puramente normativos. Este capítulo argumenta que el camino viable no reside en dotar a las máquinas de *conciencia fenomenológica (consciousness)*, sino de *capacidad operativa (awareness)* para reconocer y actuar conforme a los valores humanos. Partiendo de la distinción filosófica crítica entre ambos conceptos, se defiende que es posible y necesario construir agentes *value-aware* (conscientes de valores) sin atribuirles un estatus moral inherente. Para ello, se presenta un marco de ingeniería centrado en la Incrustación de Valores *(Value Embedding)*, que supera la rigidez del Diseño Sensible a Valores al internalizar representaciones computacionales de los valores, permitiendo una adaptación autónoma y contextual.

El núcleo de la propuesta es la arquitectura Reflective Global Neuronal Workspace (RGNW), inspirada en la neurociencia cognitiva y estructurada en tres niveles (reactivo, deliberativo y reflectivo). Este modelo no solo ejecuta comportamientos, sino que reflexiona sobre ellos a la luz de un sistema de valores, gestionando la coherencia entre preferencias axiológicas, normas derivadas y acciones. La RGNW dota al agente de una teoría de la mente artificial que le permite inferir los valores y estados mentales de sus interlocutores, y lo posiciona como un participante activo en la negociación social de valores. Así, el capítulo trasciende la visión de la IA como mero ejecutor de reglas, presentándola como un ente capaz de comprender, adaptarse y contribuir al dinámico ecosistema moral humano. La viabilidad del enfoque se demuestra con aplicaciones prácticas en análisis de redes sociales, robótica asistencial y herramientas de apoyo a la decisión clínica, señalando un camino pragmático y urgente hacia una IA socialmente responsable.

Palabras clave: conciencia de valores, sistemas *value-aware*, RGNW, inteligencia artificial, ingeniería.

* Este trabajo ha sido financiado por el proyecto Europeo VALAWAI (#101070930) y por el proyecto nacional VAE (#TED2021-131295B-C31).

1. INTRODUCCIÓN: MÁS ALLÁ DEL DISCURSO APOCALÍPTICO

En el debate contemporáneo sobre la inteligencia artificial (IA), resulta frecuente encontrar un discurso catastrofista, promovido en ocasiones por figuras prominentes y grandes corporaciones, que predica el advenimiento de una superinteligencia que supondría el fin de la humanidad (Bostrom, 2014). Este relato, a menudo revestido de un velo seudofilosófico, se sustenta en la hipótesis de que la IA alcanzará un estado de conciencia (*consciousness*) que la llevará a concluir que el mundo funcionaría mejor sin seres humanos. Es imperioso combatir esta visión, no solo por sus posibles intereses económicos subyacentes, y la desinformación que provoca en la sociedad, sino también por su frágil base conceptual.

El objetivo de este capítulo es doble. En primer lugar, se argumentará, desde planteamientos filosóficos actuales, que la IA no alcanzará la conciencia en los términos en que se discute para los seres vivos. En segundo lugar, y este es el núcleo de nuestra propuesta, se defenderá que la falta de conciencia no es un impedimento para desarrollar sistemas de IA moralmente responsables. Es perfectamente posible, y deseable, construir agentes artificiales cuyo comportamiento sea éticamente aceptable y esté alineado con los valores humanos, sin necesidad de atribuirles un estatus moral. Este enfoque se enmarca en proyectos de investigación europeos, como los financiados en el programa EIC Pathfinder "Awareness Inside", https://awarenessinside.eu), que buscan conectar la neurociencia y la IA para explorar la definición y la ingeniería de un cierto tipo de "atención consciente" y de los valores en sistemas computacionales.

Para ello, nos centraremos en el concepto de *awareness* (conciencia en el sentido de "darse cuenta" o "ser consciente de"), distinto de *consciousness* (conciencia plena o fenomenológica), y su aplicación a los valores humanos. Abordaremos la diferencia entre las aproximaciones del Diseño Sensible a Valores (*Value Sensitive Design*), que no usaremos, y la de la Incrustación de Valores (*Value Embedding*). Detallaremos una arquitectura computacional inspirada en el "Global Neuronal Workspace" (GNW) (Baars and Alonzi, 2018) para lograr sistemas *value-aware*. Finalmente, ilustraremos la viabilidad de este marco con aplicaciones concretas en redes sociales, robótica social y protocolos médicos.

2. DESENTRAÑANDO LA CONCIENCIA: *AWARENESS* VS. *CONSCIOUSNESS* Y EL ESTATUS MORAL

Un primer paso crucial es la distinción terminológica y conceptual. En inglés, y por extensión en la literatura especializada, se manejan dos términos que en castellano se traducen comúnmente como "conciencia": *consciousness* y *awareness*.

- *Consciousness*. Hace referencia a la conciencia fenomenológica, el estado subjetivo de experimentar, de tener una experiencia en primera persona. Está intrínsecamente ligado al concepto de *qualia* (las cualidades subjetivas de las experiencias, como el rojo de una rosa o el dolor de una herida). La *consciousness* es un campo de estudio en el que neurocientíficos y filósofos no han alcanzado un consenso, con decenas de definiciones en pugna (Chalmers, 1996).

- *Awareness*. Se refiere a un concepto más operacional y menos enigmático. Es la capacidad de "darse cuenta" de algo, de tener acceso a una representación de un elemento (un objeto, un hecho, un valor) y poder procesarlo cognitivamente. En el contexto de la IA, buscamos construir sistemas que sean *aware* de aspectos específicos, como la presencia de valores éticos en su entorno.

Esta distinción es fundamental para desacoplar la discusión sobre el comportamiento moral de la IA de la espinosa cuestión de su posible conciencia. Nuestra propuesta se centra exclusivamente en el *awareness*, no en la *consciousness*.

La cuestión del estatus moral de las entidades no humanas es central en la ética filosófica. Según el investigador Ali Ladac (Ladak, 2024), una entidad tiene categoría moral (*Moral standing* en su versión en inglés) si ella, o sus intereses, importan de manera intrínseca, al menos hasta cierto grado, en la evaluación moral de las acciones. Dos tipos de entidades se ajustan a esta definición:

- *Entidades sintientes* (i.e. con capacidad de sentir): la sintiencia es la capacidad de una entidad de tener experiencias positivas o negativas de manera intrínseca. Es decir, consideramos aquí a entidades para las cuales una experiencia puede ser per se buena o mala. Esta capacidad se asocia típicamente a los seres vivos.

- *Entidades conscientes:* como se ha definido, aquellas con capacidad de tener experiencias subjetivas.

Bajo esta visión, ejemplos de entidades que poseen categoría moral serían los humanos conscientes (incluso bajo los efectos de drogas, por su potencialidad de recuperación) y, de manera cada vez más aceptada, los animales. De hecho, legislaciones como la del Reino Unido han reconocido formalmente que los animales domésticos (en 2021) y los pulpos (en 2024) son *sentient*, un primer paso hacia la concesión de derechos. Una roca, una nación, serían ejemplos de entidades sin categoría moral.

El filósofo David Chalmers (Chalmers, 1996) explora esta idea mediante un experimento mental basado en el dilema del tranvía. En su versión, se debe decidir entre desviar un

tren para matar a un humano consciente o a un número **n** de zombies filosóficos (entes funcionalmente idénticos a humanos pero carentes de conciencia y de capacidad de sentir). La mayoría de las personas elige salvar al humano consciente (aunque cuando la **n** es muy grande algunos humanos empiezan a dudar). Esto ilustra que la sintiencia y la conciencia son criterios fundamentales para la atribución de categoría moral.

Aplicando este marco a la IA, se debe concluir que, actualmente, los sistemas de IA no tienen una categoría moral. No son sintientes (no tienen experiencias intrínsecamente positivas o negativas) y no poseen intereses intrínsecos. A una IA le es indiferente si está ejecutándose o no, si consigue o no un objetivo y no tiene experiencias subjetivas de ningún tipo. Por lo tanto, carece de los requisitos necesarios para ser considerada una entidad moral, lo que implica, por ejemplo, que no existe un problema ético en "apagarla".

Cabe mencionar que según otros autores se pueden usar criterios alternativos para atribuir categoría moral, como la posesión de capacidades cognitivas (agencia, inteligencia, racionalidad), la capacidad de mantener relaciones sociales o la performatividad (comportarse *como si* tuviera categoría moral). Con estos criterios nos adentramos en terrenos más resbaladizos en los que la atribución de categoría moral a sistemas de IA avanzados podría ser objeto de debate. No obstante, para los fines de este capítulo, partimos de la premisa de que es posible y deseable construir sistemas *aware* de los valores sin necesidad de atribuirles una categoría moral.

3. LA INGENIERÍA DE SISTEMAS *VALUE-AWARE*

La propuesta central de este capítulo es la ingeniería de sistemas de IA que sean *aware* de los valores humanos. La motivación es clara: si un sistema de IA se despliega en una sociedad con unos valores determinados (por ejemplo, en Japón, España o Estados Unidos), su comportamiento debe adaptarse a respetar el marco axiológico de cada contexto.

3.1. Diseño sensible a los valores vs. incrustamiento de valores

Existen dos aproximaciones principales para incorporar valores en la IA:

1. *Value Sensitive Design* (VSD). En este enfoque los valores se identifican y se presentan explícitamente a los ingenieros, quienes los incorporan en las fases de diseño y prueba del sistema. El VSD es valioso, pero puede ser rígido, ya que requiere volver a la mesa de diseño cada vez que el contexto de aplicación cambia.

2. *Value Embedding* (Incrustación de valores). Esta es la aproximación que defendemos. Los valores se representan computacionalmente de manera que el sistema los tiene "embebidos" o internalizados. Esto le permite adaptarse autónomamente a situaciones nuevas, tomando decisiones alineadas con esos valores sin necesidad de una reingeniería constante.

Una definición informal de un sistema *value-aware* sería: un sistema de IA capaz de capturar un sistema de valores dado (como la Teoría de los fundamentos morales[1] o el sistema de Schwartz[2]), entenderlo, respetarlo en sus acciones y explicar su comportamiento (y el de otros) en términos de este conjunto de valores. Los beneficios de tales sistemas se presentan en la tabla 1.

TABLA 1

BENEFICIOS DE UNA IA *VALUE-AWARE*

Función	*Descripción*
Monitorización	Verificar si las acciones propias o ajenas respetan un conjunto de valores.
Asesoramiento	Recomendar a un humano acciones compatibles con ciertos valores, justificándolas.
Certificación	Avalar si un sistema o proceso cumple con los valores establecidos.
Aprendizaje	Identificar e inferir los valores y normas implícitas en un grupo humano a partir de su comportamiento.
Autonomía guiada	Dotar de autonomía a los sistemas para la toma de decisiones, con garantías de respeto a los valores.

3.2. Una definición computacional de *awareness*

Para que el concepto de *awareness* sea operativo, necesitamos una definición computacional. En el marco del proyecto VALAWAI (https://valawai.eu/), hemos establecido que un sistema puede considerarse *aware* de un elemento *X* en un contexto *C* si cumple con tres condiciones:

1. *Acceso a una representación:* el sistema tiene acceso a una representación computacional de *X* en el contexto *C*.

2. *Coherencia:* dicha representación es coherente, sólida (*sound*) y carente de ambigüedades críticas.

[1] *Teoría de los fundamentos morales* (MFT) (Haidt y Graham, 2007): postula cinco (luego seis) pares de fundamentos intuitivos que subyacen a los juicios morales: Cuidado/Daño, Justicia/Engaño, Lealtad/Traición, Autoridad/Subversión y Santidad/Degradación.

[2] *Teoría de los valores humanos básicos de Schwartz* (Schwartz, 2012): propone una estructura circular de diez valores motivacionales universales (por ejemplo, Autodirección, Seguridad, Logro, Benevolencia) que se relacionan por compatibilidad o conflicto.

3. *Expresabilidad*: la representación puede expresarse en términos comprensibles para los humanos, permitiendo el diálogo y la justificación de las acciones.

Esta definición traslada el problema filosófico a uno de ingeniería: cómo representar, hacer coherente y expresar un valor.

3.3. El triángulo valores-normas-comportamiento: el *moral stance computacional*

En nuestra aproximación, la representación de un valor *V* se basa en dos pilares:

1. *Preferencias sobre estados del mundo.* Un valor se entiende como una preferencia sobre los posibles estados del mundo. Por ejemplo, el valor "igualdad de género" prefiere estados del mundo donde mujeres y hombres reciben el mismo salario por el mismo trabajo frente a estados donde esto no ocurre.

2. *Normas alineadas.* Las normas son la traducción operacional y deontológica de los valores. Prescriben el comportamiento correcto (por ejemplo, "debe pagarse el mismo salario por el mismo trabajo").

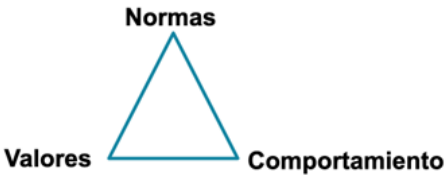
La interacción entre valores, normas y comportamiento forma un triángulo que define el *moral stance* (posicionamiento moral) de un sistema. Permite verificar:

- Si un comportamiento está alineado con un valor.
- Si un comportamiento respeta una norma.
- Si las normas y los valores ordenan los "mundos posibles" de manera consistente.

Este "alineamiento expresable" es la base práctica para construir sistemas *value-aware*.

FIGURA 1

TRIÁNGULO DE POSICIONAMIENTO MORAL



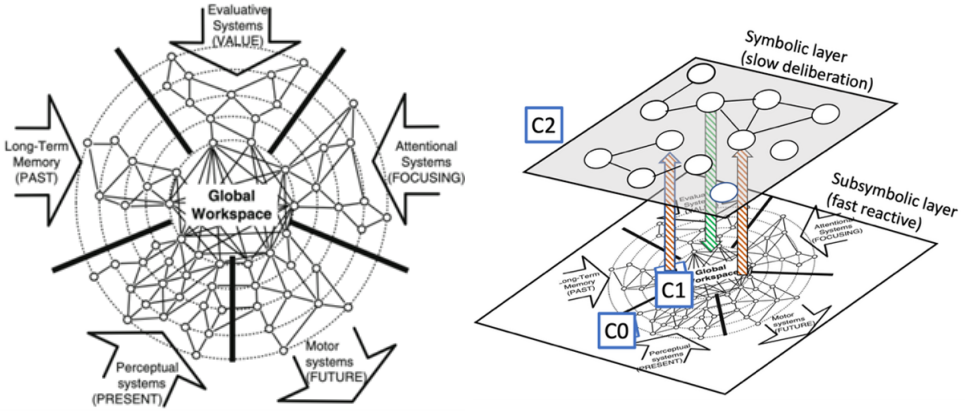
Fuente: Elaboración propia.

3.4. Una arquitectura para la conciencia de valores: el modelo RGNW

Para materializar estos principios, hemos desarrollado una arquitectura inspirada en el *Global Neuronal Workspace (GNW)* de la neurociencia (Baars y Alonzi, 2018), extendido con una capa reflectiva, dando lugar al modelo *Reflective Global Neuronal Workspace (RGNW)* (ver figura 2). Esta arquitectura se inspira también en cierta medida en el modelo dual de Kahneman (pensamiento rápido y lento) (Kahneman, 2011) y en la arquitectura de Brooks (Brooks, 1991). Consta de tres niveles:

FIGURA 2

ESQUEMA DEL MODELO GNW SEGÚN (DEHAENE *ET AL.*, 1998) (IZQUIERDA).
DIAGRAMA DE LA ARQUITECTURA RGNW USADA EN EL PROYECTO VALAWAI (DERECHA)



Fuente: Elaboración propia.

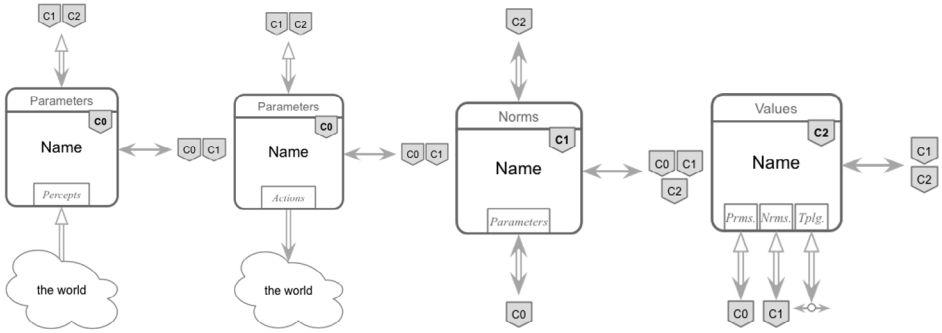
- Nivel C0 (subsimbólico/reactivo). Este nivel contiene componentes modulares que interactúan directamente con el mundo a través de sensores y actuadores. Su funcionamiento es rápido y basado en estímulos. *Ejemplo: Un módulo de* "social sensing" *que analiza tweets en tiempo real para detectar mensajes de odio.*

- Nivel C1 (deliberativo). Aquí la información de los módulos C0 se integra y se procesa de manera más simbólica. Se activan diferentes módulos en función del contexto y la percepción para tomar decisiones que se envían a los actuadores.

- Nivel C2 (reflectivo). Es el nivel superior, que alberga la representación simbólica de los valores. Su función es "reflejar" o monitorizar lo que ocurre en C0 y C1. Basándose en los valores, el nivel C2 puede *modificar la topología de la*

FIGURA 3

ESQUEMAS DE LOS COMPONENTES C0 (SENSOR Y ACTUADOR), C1 Y C2



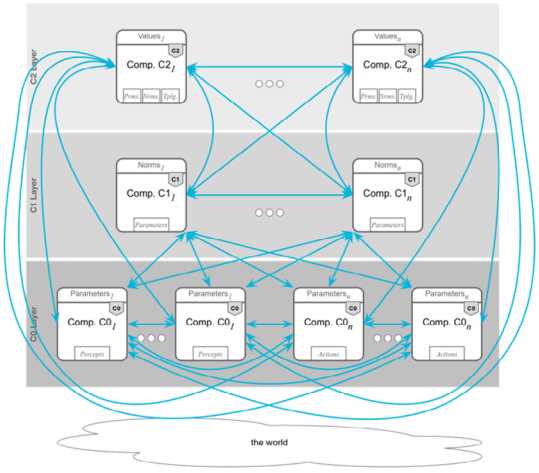
Fuente: Elaboración propia.

arquitectura, alterando qué módulos de C1 se comunican entre sí o con C0. Esto permite una adaptación profunda del comportamiento para alinearlo con los valores, sin reescribir el código de los módulos de nivel inferior.

Esta arquitectura proporciona un marco flexible y potente para implementar sistemas que no solo actúan, sino que también reflexionan sobre sus acciones a la luz de un sistema de valores internalizado. Como se ilustra en las figuras 4 y 5, el modelo RGNW

FIGURA 4

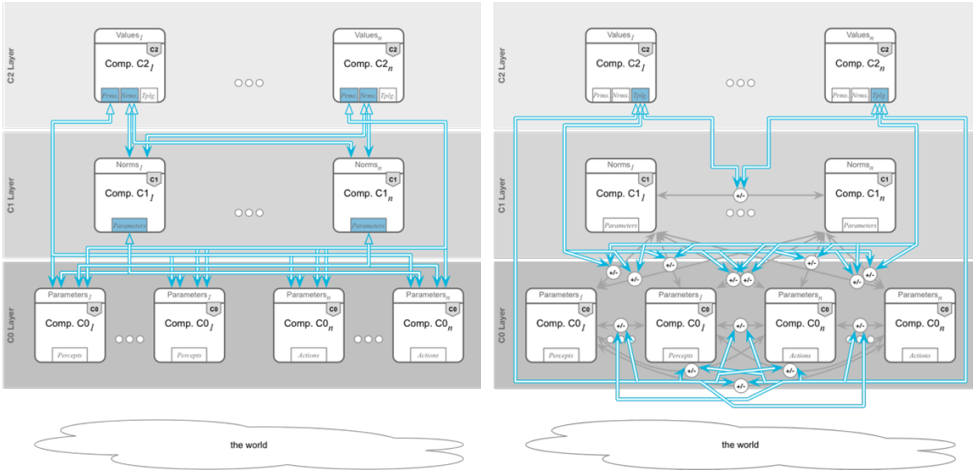
FLUJO DE DATOS EN LA ARQUITECTURA RGNW



Fuente: Elaboración propia.

FIGURA 5

FLUJO DE CONTROL (IZQUIERDA) Y DE LA TOPOLOGÍA (DERECHA) DE LA ARQUITECTURA RGNW



Fuente: Elaboración propia.

combina influencias de la arquitectura GNW y de la arquitectura de subsunción de Brooks, definiendo dos flujos esenciales:

1. *Un flujo de datos*, que garantiza la *representación, el acceso y la expresión* de la información, materializando así los puntos 1 y 3 de la definición operativa de *awareness*.

2. *Un flujo de control*, que, mediante la gestión dinámica de la *topología* de módulos, asegura la *coherencia y solidez* de las representaciones, cumpliendo con el punto 2 de dicha definición.

De este modo, la arquitectura traduce formalmente los fundamentos teóricos a un mecanismo computacional viable para la conciencia de valores.

Desde el punto de vista computacional, la arquitectura funciona como un sistema plenamente concurrente en el que todos los módulos están activos de forma continua e intercambian información. En el nivel C0, los módulos reactivos y subsimbólicos observan constantemente el entorno y procesan los estímulos entrantes, generando elaboraciones parciales que se transmiten a los niveles superiores. Cuando esta información impacta en el nivel C2, los módulos reflectivos la evalúan a la luz del sistema de valores internalizado y pueden emitir directrices normativas hacia C1. De manera crucial, este proceso reflectivo incluye una reconfiguración dinámica de la topología de la arqui-

tectura: el nivel C2 puede activar, inhibir o redirigir las conexiones entre los módulos de C1 y C0. Al modificar estos patrones de interacción, el sistema determina cómo se desarrollarán los cómputos posteriores, introduciendo un mecanismo de metacognición sobre su propio funcionamiento. Este ciclo continuo de flujo de información, adaptación topológica y cómputo concurrente acaba repercutiendo en el nivel C0, donde determinados módulos ejecutan acciones sobre el entorno, ya sean digitales o físicas, cerrando así el bucle percepción–deliberación–acción.

En este sentido, la arquitectura propuesta se aparta deliberadamente del patrón dominante en los sistemas generativos actuales, en los que el comportamiento global queda fundamentalmente determinado por un único algoritmo de aprendizaje automático entrenado a partir de grandes volúmenes de datos de entrada y salida. En lugar de basarse en una función de predicción monolítica, el modelo articula el comportamiento del sistema a través de la interacción dinámica de módulos especializados y de un mecanismo explícito de control y reflexión basado en valores. Esto no excluye, sin embargo, la incorporación de procesos de aprendizaje: el sistema podría integrar bucles de retroalimentación que evalúen las consecuencias de sus acciones en el entorno, permitiendo ajustar tanto las decisiones sobre la topología de la arquitectura como los procesos internos de los propios módulos. De este modo, el aprendizaje no sustituiría a la estructura deliberativa y reflexiva, sino que la complementaría, abriendo una línea de trabajo futura especialmente prometedora para el desarrollo de sistemas capaces de adaptarse de manera progresiva y responsable a contextos cambiantes.

3.5. Aplicaciones prácticas y casos de uso

El marco teórico y la arquitectura RGNW se han validado en varios casos de uso.

3.5.1. Análisis de valores en redes sociales

En colaboración con Sony en París, se ha desarrollado un sistema para analizar la conversación pública en Twitter (actualmente X) utilizando la Teoría de los fundamentos morales (MFT) de Haidt y Graham (2007), que postula cinco pares de valores (Cuidado/Daño, Justicia/Engaño, Lealtad/Traición, Autoridad/Subversión y Santidad/Degradación). A nivel computacional, la arquitectura RGNW se utiliza para estructurar el procesamiento de la información en tres niveles: los módulos del nivel C0 reciben y codifican los datos de entrada (los *tweets* de los usuarios); los módulos C1 procesan y extraen características relevantes, como patrones de comportamiento y expresiones de valores; y el nivel C2 evalúa estas representaciones frente a los valores internalizados, ajustando la interpretación de los *tweets* y corrigiendo, si es necesario, las clasificaciones generadas en C1. Cada usuario se posiciona en un pentágono de valores a partir de esta evaluación, y los resultados muestran una clara distribución ideológica: los perfiles

progresistas se concentran en "Cuidado" y "Justicia", mientras que los conservadores se orientan hacia "Autoridad" y "Lealtad". Esta aplicación, relativamente sencilla dentro del marco RGNW, permite automatizar el mapeo y comprensión de la ecología moral de una red social, manteniendo la coherencia entre la categorización de los *tweets* y los principios y valores definidos en C2 y demostrando la utilidad de la arquitectura para tareas de análisis ético a escala.

3.5.2. Robótica social para el cuidado de personas mayores

En la Universidad de Gante se ha desarrollado un robot social destinado a la interacción en entornos domiciliarios con personas mayores. Equipado con modelos de lenguaje grande (LLMs) y la arquitectura RGNW, el robot mantiene conversaciones con los usuarios respetando valores fundamentales como la dignidad y la privacidad. Los módulos del nivel C0 procesan continuamente las percepciones del entorno y de la interacción con el usuario, mientras que los módulos C1 integran esta información y generan posibles respuestas o acciones. El nivel C2 evalúa estas acciones frente a los valores internalizados y ajusta dinámicamente la topología de comunicación entre los módulos, asegurando que solo las respuestas coherentes con estos valores lleguen a C0. Gracias a este mecanismo, el sistema puede percibir cambios en las temáticas de la conversación—por ejemplo, de hobbies y rutinas a recuerdos personales o cuestiones de salud—indicativos de la creación de un vínculo de confianza. Además, el nivel C2 permite al robot adaptar su tono y contenido, modulando la interacción de manera ética y beneficiosa para el usuario, garantizando que la asistencia no solo sea funcional sino también respetuosa de los valores de la persona.

3.5.3. Brújula de valores en la asistencia médica

En el Hospital del Mar y el Hospital San Pablo, ambos de Barcelona, se ha desarrollado una "brújula de valores" para apoyar la toma de decisiones clínicas. El sistema, basado en los cuatro principios de la bioética[3] analiza el estado de un paciente y las acciones clínicas propuestas. Proporciona un *feedback* al médico, indicando en qué medida cada opción respeta o vulnera estos principios. Esto no sustituye al clínico, sino que le ofrece una herramienta de reflexión para mejorar su práctica, asegurando que las decisiones técnicas también sean éticas.

Desde un punto de vista computacional, la "brújula de valores" implementa un flujo de información jerárquico y concurrente entre los niveles de la arquitectura RGNW. La información que llega a C0 corresponde a las observaciones sobre el estado del paciente, que

[3] *Principialismo bioético* (Beauchamp y Childress, 2013): marco compuesto por cuatro principios prima facie para la deliberación ética en medicina: Beneficencia (hacer el bien), No maleficencia (no hacer daño), Autonomía (respetar la autodeterminación) y Justicia (equidad en la distribución de beneficios y cargas).

son procesadas por los módulos internos en C1 encargados de analizar datos, predecir posibles evoluciones clínicas y evaluar qué acciones son factibles. Esta información se usa para integrar y deliberar sobre las posibles acciones. El nivel C2 interviene de manera reflexiva, evaluando cada acción prevista frente a los valores internalizados (los principios de la bioética). A partir de esta evaluación, C2 puede inhibir o modificar las conexiones entre los módulos de C1, de modo que solo aquellas acciones coherentes con los valores éticos lleguen a C0 y sean finalmente sugeridas a los clínicos. Este mecanismo permite un control metacognitivo sobre la arquitectura, garantizando que las decisiones técnicas también cumplan criterios éticos antes de ser comunicadas al profesional sanitario.

3.6. Dilemas éticos y la importancia del contexto social

La ingeniería de valores que proponemos no puede eludir la complejidad de los dilemas éticos. Experimentos como el del "tranvía" y sus variantes (como la de empujar a una persona gruesa desde un puente para salvar a cinco) demuestran que la toma de decisiones morales humanas no se rige por una simple utilidad. La *causación* (si mi acción es directa o indirecta) y los vínculos emocionales (elegir entre un desconocido y un hijo) alteran profundamente el juicio.

Un experimento revelador con psicópatas de la cárcel de Lleida ilustra esta diversidad. Se planteó a los participantes si matarían a una persona sana pero molesta para utilizar sus órganos y salvar a cinco pacientes. Mientras que la mayoría de la población se opone, alrededor del 65 % de los psicópatas aceptaría la opción utilitarista, debido a su falta de empatía y remordimiento.

Este caso subraya una conclusión crucial: *los aspectos éticos de la IA no pueden ser decididos unilateralmente por ingenieros en Silicon Valley, Palo Alto o cualquier otro centro tecnológico*. Deben ser el resultado de un consenso social que involucre a los colectivos que van a utilizar y a ser afectados por la tecnología. La ética, como reflexión filosófica sobre las normas, a menudo entra en conflicto con la moral (las normas concretas de una sociedad), tal y como ilustra el conflicto de Antígona entre su obligación familiar (enterrar a su hermano) y la ley de la ciudad (prohibir el entierro a un traidor). La IA debe ser capaz de navegar en estas tensiones, para lo cual es fundamental que sus sistemas de valores sean transparentes, debatibles y adaptables a los contextos culturales y situacionales.

4. EL PAPEL DE LA TEORÍA DE LA MENTE EN LA INTERACCIÓN SOCIAL Y SU IMPLEMENTACIÓN EN LA RGNW

Para que un sistema de IA sea genuinamente *value-aware* y pueda interactuar de forma ética y efectiva en contextos sociales, debe ir más allá del procesamiento de reglas

explícitas de interacción, de protocolos rígidos. Es necesario que desarrolle una capacidad operativa de *Teoría de la mente (ToM)*. En psicología y ciencias cognitivas, la ToM se refiere a la habilidad de atribuir estados mentales (creencias, deseos, intenciones, conocimientos, emociones) a otros, y de comprender que estos pueden diferir de los propios. Esta capacidad es fundamental para predecir, interpretar e influir en el comportamiento ajeno, y es un pilar de la competencia social.

La arquitectura *RGNW* proporciona un marco natural y estructurado para implementar una ToM artificial, esencial para una conciencia de valores socialmente situada. Su implementación se distribuye a lo largo de sus tres niveles:

1. En el Nivel C0 (reactivo). Módulos especializados de "percepción social" procesan señales en tiempo real (análisis de lenguaje natural, tono de voz, expresiones faciales, postura) para generar inferencias básicas sobre estados emocionales o intenciones inmediatas. Por ejemplo, un módulo puede detectar frustración en la sintaxis o el léxico utilizado por un usuario.

2. En el Nivel C1 (deliberativo). Estas señales se integran para construir *modelos de agente.* Aquí, el sistema forma y actualiza representaciones simbólicas o subsimbólicas sobre los estados mentales atribuidos a cada interlocutor (por ejemplo, "el usuario *cree* que no he entendido su instrucción" o "el usuario *desea* terminar la conversación"). Este nivel permite el razonamiento sobre cómo las acciones propias afectarán estos estados mentales.

3. En el Nivel C2 (reflectivo/valores). Este es el componente crucial que diferencia una ToM meramente funcional de una *ToM guiada por valores*. El sistema de valores alojado en C2 actúa como un filtro y un guía para la ToM:

    - *Prioriza* qué estados mentales de otros son moralmente relevantes (por ejemplo, dar mayor peso a detectar incomodidad o confusión que a simple aburrimiento).

    - *Evalúa* las posibles acciones deliberadas en C1 no solo por su eficacia instrumental, sino por su *impacto en el bienestar, la autonomía o la dignidad del otro,* según los valores.

    - *Modifica la topología* para dirigir la atención (flujo de datos) de los módulos C1 hacia los aspectos de la interacción que son axiológicamente significativos.

De este modo, la RGNW logra que la ToM no sea un módulo aislado, sino una *capacidad permeada por valores*. Un robot social con esta arquitectura no solo inferirá que una persona mayor está repitiendo una historia, sino que, a la luz del valor de la *dignidad*, su

nivel C2 guiará la interacción para evitar reacciones que puedan hacer sentir al usuario ignorado o menospreciado. En una negociación automática, el sistema podría utilizar su ToM para identificar cuándo la contraparte actúa por necesidad (estado mental) versus por ambición, ajustando sus propuestas para alinearse con el valor de la *justicia*.

Por lo tanto, la integración de una Teoría de la mente dentro de la arquitectura RGNW es un paso indispensable para trascender de sistemas que *conocen* valores abstractos a sistemas que los *aplican* en dinámicas sociales complejas, interactuando no con agentes genéricos, sino con entidades cuyos estados mentales deben ser comprendidos y respetados.

5. NEGOCIACIÓN DE VALORES Y ESTRATEGIAS BASADAS EN LA RGNW

La arquitectura RGNW no solo permite a un agente actuar conforme a un sistema de valores estático, sino que lo dota de la capacidad fundamental para participar en la negociación dinámica de valores que caracteriza a las comunidades humanas. Este proceso implica dos dimensiones entrelazadas: 1) la observación y modelado de cómo las comunidades definen y armonizan valores, y 2) la participación estratégica en diálogos o interacciones donde los valores de los participantes pueden converger o entrar en conflicto.

5.1. Modelado de la negociación social de valores

Las sociedades humanas no operan con un conjunto de valores fijo e inmutable; estos son continuamente puestos a prueba, reinterpretados y negociados a través del discurso público, las normas legales y las interacciones cotidianas. Un sistema *value-aware* debe poder observar y entender este proceso. La RGNW puede facilitar esto mediante:

- Aprendizaje e inferencia (Nivel C1). El sistema puede analizar grandes corpus de interacciones (debates parlamentarios, hilos de redes sociales, resoluciones judiciales) para identificar cómo ciertos valores (por ejemplo, "privacidad" vs. "seguridad") son invocados, ponderados y ajustados en distintos contextos. Podemos desarrollar módulos que detecten patrones de *trade-offs* y compromisos.

- Representación de consensos y conflictos (Nivel C2). El nivel reflectivo puede mantener representaciones de los "espacios de valores" de una comunidad, no como puntos fijos, sino como, por ejemplo, distribuciones de probabilidad. Puede modelar que, en una comunidad dada, el valor "libertad individual" tiene un peso alto pero está en constante conflicto con el valor "responsabilidad colectiva".

5.2. Estrategias de negociación basadas en valores

Cuando un agente con RGNW se encuentra en un escenario de negociación (cooperativa o competitiva) con humanos u otros agentes, su arquitectura le permite desarrollar estrategias sofisticadas que van más allá de la mera maximización de utilidad económica clásica. En lugar de optimizar exclusivamente un criterio utilitarista, el agente considera un conjunto de valores éticos y normativos que pueden guiar sus decisiones incluso cuando impliquen aceptar pequeñas pérdidas económicas.

La clave reside en utilizar un modelo de la Teoría de la mente (ToM) para inferir el sistema de valores del interlocutor. Esta información permite al agente alinear estratégicamente sus acciones, es decir, elegir comportamientos que no solo buscan un beneficio material, sino que también respetan o satisfacen los valores del otro agente o humano. La alineación estratégica implica seleccionar acciones que maximicen simultáneamente la utilidad clásica y la satisfacción de los valores éticos, lo que puede incluir compromisos de carácter deontológico, promoción de la cooperación y respeto a normas sociales, asegurando que la negociación sea efectiva y socialmente aceptable.

1. Detección de alineamiento y fricción axiológica. El sistema, a través de sus módulos en C0 y C1, puede analizar las propuestas y reacciones del interlocutor. A nivel C2 se puede evaluar continuamente: ¿los valores que motivan mi propuesta (por ejemplo, "eficiencia") están alineados con los valores que parecen motivar al otro (por ejemplo, "estabilidad" o "equidad")? Y de esa manera identificar tanto *valores de acuerdo* como *valores en conflicto.*

2. Generación de opciones y argumentación guiada por valores. Basándose en esta evaluación, el nivel C1/C2 puede generar no solo opciones que maximizan un objetivo, sino opciones que pueden ser *renmarcadas* o *justificadas* en términos de los valores del interlocutor.

    - Si hay *valores compartidos*, la estrategia será *amplificarlos.* "Ambas propuestas buscan la justicia, la mía lo hace priorizando la equidad a largo plazo".

    - Si hay *conflicto de valores*, la estrategia puede ser:

        - *Crear puentes.* Encontrar un valor de orden superior compartido ("Ambos valoramos el bienestar de la comunidad, exploremos cómo balancear libertad y seguridad para lograrlo").

        - *Ofrecer compensación axiológica.* Reconocer explícitamente el sacrificio de un valor y compensarlo en otro ámbito ("Esta opción prioriza la eficiencia que usted necesita, y propongo una cláusula de revisión que proteja la transparencia que yo valoro").

3. Adaptación dinámica de la topología (Función clave del C2). Durante la negociación, el nivel C2 puede *reconfigurar la prioridad de los módulos deliberativos* en C1. Por ejemplo:

   - Si detecta que el interlocutor responde positivamente a argumentos basados en "lealtad", puede potenciar los módulos que generan propuestas que enfaticen compromisos a largo plazo y relaciones estables.

   - Si percibe que la negociación está estancada por un conflicto de valores irreconciliable, puede activar módulos de "creatividad" para buscar opciones innovadoras que satisfagan los intereses subyacentes de ambos sin violar sus valores esenciales.

5.3. Ejemplo ilustrativo: negociación automatizada de un contrato

Consideremos un agente RGNW que negocia un acuerdo de colaboración entre una universidad y una empresa.

Valores del agente (Universidad). Acceso abierto al conocimiento, independencia académica.

Valores inferidos de la contraparte (Empresa). Propiedad intelectual, rentabilidad.

Proceso RGNW:

1. *C1 (Teoría de la mente–ToM).* El agente analiza la conducta de la contraparte y otros datos contextuales (por ejemplo, insistencia en cláusulas de confidencialidad) para inferir que estos comportamientos reflejan el valor "propiedad intelectual".

2. *C2 (reflexión).* Evalúa los conflictos entre los valores propios y los de la contraparte ("acceso abierto" vs. "propiedad intelectual") y busca un valor compartido superior: "innovación".

3. *C2 (modulación).* Ajusta los módulos de generación de propuestas en C1 para producir opciones que maximicen un *criterio híbrido*:

   - Satisfacción de los valores propios del agente (universidad).

   - Reconocimiento parcial de los valores de la contraparte (empresa), aunque esto implique cierta pérdida económica.

4. *Salida:* El agente propone una "ventana de embargo": los resultados son confidenciales durante 12 meses (respetando la propiedad intelectual) y luego se hacen de acceso abierto (respetando la misión universitaria). La propuesta argumenta que este modelo acelera la innovación, beneficiando a ambas partes.

La propuesta del agente no busca maximizar exclusivamente la utilidad económica propia, sino encontrar un punto intermedio que equilibre intereses económicos y valores éticos de ambas partes. Los datos que utiliza incluyen comportamientos observables de la contraparte, reglas explícitas de valores, y contexto del escenario de negociación. El agente optimiza una función híbrida que combina beneficio económico y alineación con valores, lo que le permite aceptar pequeñas pérdidas económicas para respetar o fomentar los valores del interlocutor. De esta manera, la RGNW actúa como un participante activo en el ecosistema moral de la negociación, capaz de comprender, influir y adaptarse a los procesos sociales mediante los cuales los valores se negocian y redefinen colectivamente.

6. DISCUSIÓN

La propuesta arquitectónica del *Reflective Global Neuronal Workspace (RGNW*) se enmarca, en última instancia, en el largo esfuerzo por operacionalizar principios abstractos en sistemas computacionales. Su núcleo innovador no radica en descartar los mecanismos normativos clásicos, sino en dotarlos de una capacidad de reflexión guiada por valores. Al definir un valor como una preferencia sobre estados del mundo y generar a partir de él normas alineadas que prescriben comportamientos, la RGNW implementa *de facto* un sistema normativo dinámico y autoadaptativo. Este enfoque establece un diálogo directo con los fundamentos de las instituciones electrónicas, un marco pionero en el que las normas explícitas (protocolos, reglas de compromiso, contratos) se utilizan para estructurar, gobernar y hacer predecibles las interacciones entre agentes autónomos dentro de un espacio de acción compartido. La arquitectura RGNW puede interpretarse como la internalización y "reflexivización" de este principio institucional: su nivel C2 actúa como una suerte de institución electrónica personal y autogobernada, donde los valores constituyen los objetivos de diseño de alto nivel y el conjunto de normas derivadas (cuya aplicación se gestiona mediante la topología dinámica entre C1 y C0) opera como el mecanismo de gobierno interno. Este vínculo conecta directamente con líneas de investigación previas sobre la formalización lógica y la gestión computacional de normativas en sistemas multiagente, como las desarrolladas en el contexto de las instituciones electrónicas (D'Inverno *et al,*, 2012). La contribución distintiva de la RGNW es la adición de la capa reflectiva (C2), que permite al sistema no solo ejecutar normas, sino también evaluar continuamente su coherencia con el sistema de valores, adaptarlas contextualmente y explicitar la justificación axiológica de sus acciones. Así, se transita desde un gobierno normativo estático hacia una administración consciente y justificada de la coherencia axiológica, un paso crucial para construir agentes autóno-

mos que operen con responsabilidad en entornos sociales complejos, es decir con una multitud de agentes autónomos con objetivos e intereses a menudo contrapuestos.

Este capítulo ha defendido una tesis central: la responsabilidad moral de la inteligencia artificial se construye mediante ingeniería, no mediante la especulación sobre una conciencia inalcanzable. Al desacoplar conceptualmente la *conciencia fenomenológica* (*consciousness*) de la *capacidad operativa de darse cuenta* (*awareness*), hemos reenfocado el problema desde la metafísica hacia la computación. La propuesta de sistemas *value-aware*, materializada en la arquitectura RGNW, constituye una respuesta concreta a este desafío.

La RGNW demuestra que es factible diseñar agentes que internalicen valores como preferencias sobre estados del mundo, los traduzcan en normas operativas y monitoricen su propia coherencia axiológica a través de una capa reflectiva. Más allá de la adhesión estática a reglas, esta arquitectura permite a los agentes navegar la complejidad social: implementar una Teoría de la mente para interpretar a los demás, participar en la negociación dinámica de valores y reconfigurar su propio funcionamiento para alinearse con principios éticos en contextos cambiantes. En este sentido, la RGNW opera como una institución electrónica internalizada, un sistema normativo autogobernado y guiado por valores que evoca el trabajo fundacional en sistemas multiagente.

Las aplicaciones descritas —desde el mapeo de la ecología moral en redes sociales hasta la asistencia robótica sensible al contexto— no son meros prototipos, sino pruebas de concepto que validan el marco teórico. Evidencian que la "conciencia de valores" es un requisito técnico imprescindible para el despliegue responsable de la IA en dominios sociales.

El camino futuro no está exento de retos. La explicabilidad profunda de las decisiones guiadas por valores, la gestión de conflictos axiológicos irresolubles y el diseño de mecanismos para la evolución y aprendizaje colaborativo de valores con las comunidades humanas son fronteras de investigación abiertas. Sin embargo, el debate ya no debe centrarse en si las máquinas pueden ser sujetos morales, sino en cómo podemos dotarlas, de manera robusta y verificable, de la capacidad para saber lo que importa. La tarea urgente para ingenieros, diseñadores y científicos sociales es refinar estas herramientas, asegurando que la IA del futuro no solo sea inteligente, sino también centrada en valores, reflejando y sirviendo al pluralismo moral de la humanidad.

Referencias

Baars, B. J., & Alonzi, A. (2018). The global workspace theory. En R. J. Gennaro (Ed.), *The Routledge handbook of consciousness* (pp. 233-245). Routledge.

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7ª ed.). Oxford University Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1-3), 139–159.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences, 95*(24), 14529–14534.

D'Inverno, M., Luck, M., Noriega, P., Rodriguez-Aguilar, J. A., & Sierra, C. (2012). Communicating open systems. *Artificial Intelligence,* 186, 38–94.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Ladak, A. (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics,* 4, 213–228.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture, 2*(1).

Sobre los autores

Amparo Alonso Betanzos

Catedrática en Computación e Inteligencia Artificial en el CITIC de la Universidade da Coruña (UDC). Investigación en IA responsable y frugal. Académica correspondiente de la Real Academia de Ciencias Exactas, Físicas y Naturales desde 2023, y miembro del Consejo Asesor de Ética en Investigación del Gobierno de España.

Senén Barro

Director del CiTIUS-Universidad de Santiago de Compostela (USC). Más de 300 publicaciones en IA y más de 500 de divulgación y opinión. Rector de la USC entre 2002 y 2010. Premio Nacional de Informática José García Santesmases en 2020. Promotor-fundador de dos spin-offs: Situm Technologies e InVerbis Analytics.

Francisco Bellas

Es catedrático en la Universidade da Coruña (UDC), con una amplia trayectoria profesional en robótica autónoma, está actualmente involucrado en la IA aplicada a la educación, como investigador y gestor de proyectos, y colaborando como asesor a nivel europeo e internacional.

Holger Billhardt

Es catedrático en la Universidad Rey Juan Carlos en el área de Ciencia de la Computación e Inteligencia Artificial y ocupa el cargo de subdirector del Centro (CETINIA). Su investigación se centra en los sistemas multiagente, especialmente en la coordinación y en los aspectos éticos de estos sistemas.

Vicent J. Botti Navarro

Es catedrático de Universidad en la Universidad Politècnica de València. El profesor Botti es uno de los pioneros de la Inteligencia Artificial, Sistemas Multiagente y Tecnologías del Acuerdo. Su actividad abarca principalmente: Sistemas Multiagentes, Sistemas Inteligentes de Fabricación, Tecnologías del Acuerdo y Computación Afectiva.

José Miguel Burés

Graduado en Robótica por la USC (Premio Ángela Ruiz Robles a la excelencia académica) y máster en IA por la UIMP. Investigador predoctoral FPU centrado en el desarrollo de arquitecturas de aprendizaje federado para entornos con datos no IID, con resultados publicados en el congreso internacional ECAI 2025.

Alberto Fernández

Es catedrático de la Universidad Rey Juan Carlos (URJC), donde es miembro del Centro de Investigación CETINIA. Sus principales líneas de investigación son los sistemas multiagente abiertos, tecnologías semánticas para la representación del conocimiento (grafos de conocimiento y ontologías) y aspectos éticos de la IA.

Andrés Holgado Sánchez

Es doctorando en la Universidad Rey Juan Carlos. Cursó Matemáticas e Ingeniería Informática en la Universidad Autónoma de Madrid y un máster en Inteligencia Artificial en la Universidad Politécnica de Madrid. Desde 2023, publica sobre inteligencia artificial y valores éticos en conferencias como ICAART, ECAI y AAMAS.

Roberto Iglesias

Catedrático de Ciencias de la Computación e Inteligencia Artificial. Doctor en Física (2003). Investigador del CiTIUS y actualmente coordinador del grado de robótica de la USC. Miembro del grupo de Sistemas Inteligentes. Autor de más de 110 artículos científicos y socio fundador de la spin-off SITUM Technologies.

David Martínez Rego

Es un investigador español en aprendizaje automático y emprendedor, cofundador de DataSpartan, empresa londinense de IA y Ciencia de Datos que aborda retos complejos de Machine Learning y Big Data. Es doctor en computación e inteligencia artificial, director de investigación y contribuye académicamente mientras lidera investigación aplicada y desarrollo tecnológico.

Sascha Ossowski

Es catedrático y director del Centro de Investigación CETINIA en Inteligencia Artificial y Tecnologías de la Información de la Universidad Rey Juan Carlos. Se licenció en Informática en la Universidad de Oldenburgo (Alemania) y obtuvo el título de doctor en el Departamento de Inteligencia Artificial de la Universidad Politécnica de Madrid.

Carles Sierra

Expresidente de EurAI y director del IIIA (CSIC), es investigador en inteligencia artificial aplicada a sistemas multiagente y ética en IA. Desarrolla modelos de comportamiento social y estrategias de negociación automatizada, liderando iniciativas sobre IA responsable y alineación de valores en entornos tecnológicos.

Funcas
Caballero de Gracia, 28
28013 Madrid
Teléfono: 91 596 54 81

publica@funcas.es
www.funcas.es

ISBN 979-13-87770-17-4