GRUPO EDITORIAL DE PSICOFUNDACIÓN

Apuntes de Psicología (2026) 44(2) 77-84

Apuntes de Psicología

https://www.apuntesdepsicologia.es • ISSN 0213-3334 • eISSN 1989-6441

Colegio Oficial de Psicología de Andalucía Occidental

Universidad de Cádiz
Universidad de Córdoba
Universidad de Huelva
Universidad de Sevilla

Artículo

Inteligencia Artificial Aplicada a la Psicología: El Ejemplo de *Psypilot* Como Copiloto Terapéutico

Pablo Roca[1,2], Martin Sanchez-Pedreño[1,2], Guillermo Rodriguez-Fernandez[2], Eduardo P. Garcia del Valle[2,3] y Rosaria Maria Zangri[2,4]

[1]Universidad Villanueva, Madrid (España)
[2]Medea Lab, Madrid (España)
[3]IE Universidad, Madrid (España)
[4]UNIE Universidad, Madrid (España)

INFORMACIÓN

Recibido: 22/01/2026
Aceptado: 30/03/2026

Palabras clave:
Inteligencia artificial
Salud mental de precisión
Toma de decisiones basada en datos
Cuidado basado en la medición

RESUMEN

La Inteligencia Artificial (IA) está remodelando la práctica psicológica, planteando una encrucijada estratégica entre el despliegue de "terapeutas automatizados" y el desarrollo de "copilotos digitales" que potencien el juicio clínico sin sacrificar el vínculo profesional. Este artículo sintetiza las aplicaciones de la IA en psicología, desde la evaluación hasta la documentación, analizando sus implicaciones prácticas para la calidad asistencial. Frente a los modelos de IA sustitutiva, se presenta el caso de *Psypilot* como ejemplo de implementación de un copiloto terapéutico para potenciar la labor profesional, ilustrando cómo operacionalizar el paradigma de Salud Mental de Precisión en la práctica clínica. Finalmente, se examinan los desafíos de gobernanza y los dilemas deontológicos, proponiendo un marco ético donde la tecnología actúe como soporte a las decisiones clínicas. Se concluye que el futuro de la profesión no pasa por la resistencia tecnológica, sino por la adquisición de nuevas competencias y la alfabetización algorítmica.

Artificial Intelligence Applied to Psychology: The Example of Psypilot as a Therapeutic Co-Pilot

Keywords:
Artificial intelligence
Precision mental health
Data-driven decision-making
Measurement-based care

ABSTRACT

Artificial Intelligence (AI) is transforming psychological practice, reaching a strategic crossroads where "automated therapists" are being deployed alongside "digital co-pilots" to enhance clinical judgement without compromising the professional bond. This article summarizes the applications of AI in psychology, from assessment to documentation, analyses its practical implications for the quality of care. Unlike substitute AI models, the case study of *Psypilot* is presented as an example of implementing a therapeutic co-pilot to support professional work, demonstrating how to implement the Precision Mental Health paradigm in clinical practice. Finally, the article examines governance challenges and ethical dilemmas, proposing an ethical framework in which technology supports clinical decision-making. The conclusion is that the future of the profession lies not in technological resistance, but in acquiring new skills and becoming algorithmically literate.

Citar como: Roca, Pablo; Sanchez-Pedreño, Martin; Rodriguez-Fernandez, Guillermo; Garcia del Valle, Eduardo P. y Zangri, Rosaria Maria (2026). Inteligencia Artificial aplicada a la psicología: El ejemplo de *Psypilot* como copiloto terapéutico. *Apuntes de Psicología, 44*(2), 77-84. https://doi.org/10.70478/apuntes.psi.2026.44.09

Autor y e-mail de correspondencia: Pablo Roca, pablo.roca@medeamind.com

¿De qué Hablamos Cuando Hablamos de IA?

Para entender el alcance de las herramientas de IA en psicología, es necesario desmitificar la terminología técnica y aterrizarla en la práctica psicológica. De manera sencilla, la IA podría definirse como sistemas de software capaces de percibir el entorno, procesar la información y actuar de maneras que son (en cierta medida) orientadas a objetivos y adaptativas. Esta idea se alinea con la definición clásica de Russell y Norvig (2020) de la IA como el estudio de agentes que reciben percepciones del entorno y realizan acciones. Por tanto, la IA no es una tecnología única, sino un término "paraguas" que agrupa diversos sistemas capaces de procesar información para perseguir objetivos.

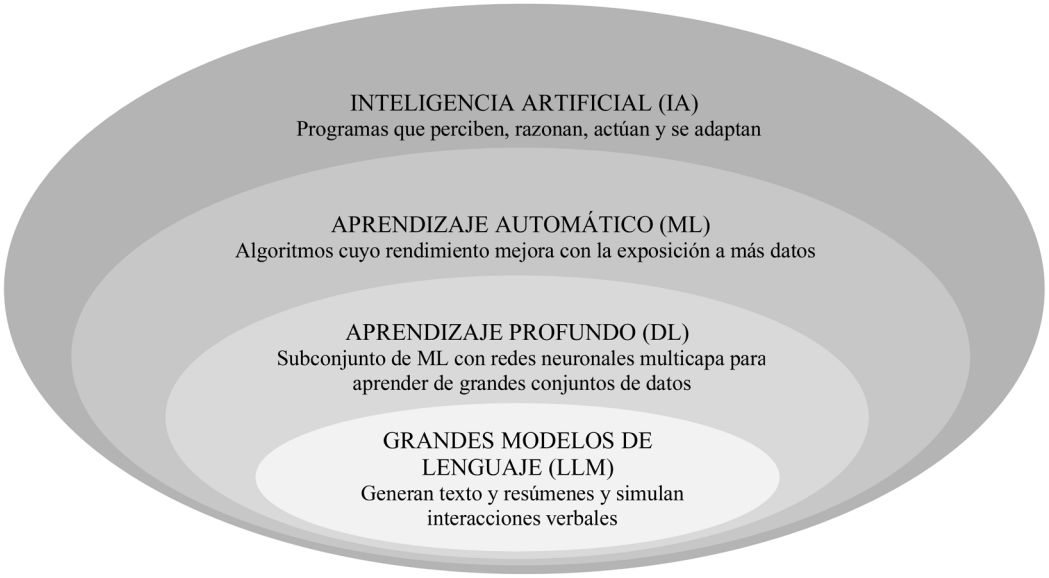
Dentro de este amplio paraguas (ver figura 1), el Aprendizaje Automático (*Machine Learning* - ML) es la disciplina más relevante para nuestra profesión. A diferencia del software tradicional, donde un humano escribe reglas fijas, los algoritmos de ML detectan patrones estadísticos en grandes volúmenes de datos. En investigación en salud mental, el ML se utiliza típicamente para aprender patrones que conectan datos de entrada (p. ej., respuestas a cuestionarios, notas clínicas, características de la voz, trazas del *smartphone*) con salidas (p. ej., diagnóstico, riesgo de recaída, probabilidad de responder a un tratamiento determinado). A su vez, dentro del subconjunto de ML, destaca el Aprendizaje Profundo (*Deep Learning* - DL), que utiliza redes neuronales de múltiples capas para aprender representaciones jerárquicas: combinaciones de características que se vuelven progresivamente más abstractas a través de las capas. El DL permite a los ordenadores entender el mundo en términos de una jerarquía de conceptos, aprendida directamente de los datos sin necesidad de especificar manualmente todas las características relevantes.

Para los psicólogos, dos variantes de IA son especialmente relevantes dado el papel fundamental de la palabra en el proceso terapéutico: el Procesamiento del Lenguaje Natural (*Natural Language Processing* - NLP), rama de la IA que permite a las máquinas analizar, comprender y extraer significado del lenguaje humano (texto o habla). En el contexto clínico, actúa como el traductor que convierte los datos cualitativos no estructurados (i.e., lo que el paciente o terapeuta dice o escribe) en datos estructurados que los algoritmos pueden procesar, permitiendo detectar marcadores lingüísticos de depresión o resumir notas de terapia. Revisiones sistemáticas muestran un rápido crecimiento del uso de NLP para procesar notas clínicas, cuestionarios, transcripciones terapéuticas y redes sociales para tareas como el diagnóstico o la predicción de riesgo (Le Glaz et al., 2021); por otro lado, los Grandes Modelos de Lenguaje (*Large Language Model* - LLM): capaces de producir texto, resúmenes, sugerencias… Estos modelos sustentan los copilotos de IA actuales y se utilizan cada vez más para redactar borradores de informes, realizar informes clínicos o simular interacciones verbales completas (Lawrence et al., 2024).

Desde la perspectiva del psicólogo, las distinciones más importantes no son entre algoritmos específicos (p. ej., *random forest vs. support vector machine*), sino entre los diferentes tipos de aprendizaje que realizan, pudiendo distinguirse cuatro tipologías clave para la psicología (Dwyer et al., 2018; Shatte et al., 2019):

1. *Aprendizaje supervisado:* Es el estándar en el diagnóstico asistido. Entrenamos al algoritmo con ejemplos etiquetados donde se conocen tanto las entradas como las salidas (p. ej., perfiles de síntomas + etiquetas diagnósticas; datos de línea base + respuesta al tratamiento). Las tareas incluyen clasificación (p. ej., deprimido vs. no deprimido; alto vs. bajo riesgo de abandono terapéutico) y regresión (p. ej., predecir severidad de síntomas, número de sesiones necesarias, días hasta la recaída).
2. *Aprendizaje no supervisado:* Aquí no damos etiquetas ni respuestas correctas al sistema, sino que le pedimos que encuentre estructuras ocultas en los datos (p. ej., agrupaciones de síntomas, subtipos dentro de un trastorno). Las tareas comunes incluyen el *clustering* (descubrir grupos de individuos similares) y la reducción de dimensionalidad (identificar un conjunto más pequeño de dimensiones latentes que resuman muchas variables). En salud mental,

Figura 1
Paraguas de Técnicas Incluidas Dentro de la IA



los métodos no supervisados se usan para explorar subtipos novedosos de depresión o psicosis, identificar dimensiones latentes de la psicopatología o segmentar poblaciones para intervenciones dirigidas.

3. *Aprendizaje semisupervisado y débilmente supervisado:* muchos conjuntos de datos en salud mental contienen pocos casos etiquetados y una cantidad mucho mayor de datos no etiquetados (p. ej., texto en bruto). Los métodos semisupervisados aprovechan ambos para mejorar el rendimiento, lo cual es especialmente relevante cuando el etiquetado clínico es costoso.
4. *Aprendizaje por refuerzo:* El algoritmo aprende por ensayo y error interactuando con un entorno para maximizar las recompensas y minimizar las penalizaciones. En psicología, se está explorando el aprendizaje por refuerzo para optimizar intervenciones adaptativas *just-in-time* (p. ej., decidir cuándo enviar un aviso o qué tipo de micro intervención ofrecer según el contexto momentáneo).

Cualquier sistema de IA es tan bueno —o tan sesgado— como los datos de los que aprende. En psicología, los modelos de IA se alimentan de cuatro fuentes principales de información:

1. *Datos psicológicos estructurados:* Es la información clásica y estandarizada, como las respuestas a cuestionarios psicométricos, códigos diagnósticos (CIE/DSM) o registros de asistencia y medicación presentes en la historia clínica electrónica.
2. *Datos textuales y lingüísticos:* Gracias al NLP y los LLM, hoy podemos procesar datos no estructurados como las notas clínicas redactadas por el profesional, transcripciones de sesiones de psicoterapia o incluso narrativas escritas por el paciente en diarios digitales o correos electrónicos.
3. *Datos de audio, vídeo e interacción:* Esta capa analiza características paralingüísticas (prosodia, tono de voz, latencia de respuesta o silencios) y señales no verbales capturadas en vídeo. Estos marcadores son especialmente relevantes para detectar incongruencias afectivas o cambios sutiles en el estado de ánimo.
4. *Fenotipado digital y datos de sensores:* Se refiere a los datos pasivos recogidos por smartphones y wearables (relojes inteligentes), como patrones de movilidad (GPS), regularidad del sueño, variabilidad de la frecuencia cardíaca o frecuencia de comunicación social.

Es crucial entender que la procedencia de estos datos determina la validez del modelo. Si los algoritmos se entrenan con muestras poco representativas, el resultado carecerá de validez ecológica para la población general, un riesgo ético central en la implementación de estas tecnologías (Lee et al., 2021). Muchos conjuntos de datos sobre representan ciertos países, sistemas sanitarios o grupos demográficos, lo que puede dar lugar a modelos que funcionan bien "sobre el papel" pero generalizan mal a poblaciones diversas del mundo real, contribuyendo así a problemas de injusticia algorítmica. Para los psicólogos, esto implica que entender de dónde provienen los datos forma parte del uso ético de la IA (Lee et al., 2021).

La Irrupción de la IA en la Atención Sanitaria

Aunque la psicología se encuentra en una fase temprana de adopción de este tipo de tecnologías, podemos anticipar su impacto observando la trayectoria de la medicina general, donde la IA ha transitado de prototipos experimentales a herramientas de uso cotidiano. Evidencias recientes demuestran que los algoritmos pueden igualar o superar el rendimiento de médicos especialistas en tareas bien definidas. Ejemplos notables incluyen sistemas que alcanzan precisión de nivel experto en la clasificación de cáncer de piel (Esteva et al., 2017) o modelos que reducen las tasas de falsos positivos y negativos en el cribado de cáncer de mama (McKinney et al., 2020).

Estos precedentes médicos ponen de relieve una serie de lecciones valiosas para la psicología: (1) cuando los problemas están bien delimitados, se dispone de grandes volúmenes de datos y los resultados pueden medirse con claridad, la IA puede igualar e incluso superar el desempeño de especialistas; (2) los despliegues más exitosos no funcionan como "cajas negras" autónomas, sino como sistemas de soporte a las decisiones clínicas (*Clinical decision support systems* o CDSS; Schaffrath et al., 2022), donde la IA no reemplaza el juicio del oncólogo o el radiólogo, sino que actúa como una segunda opinión basada en datos; y (3) se están consolidando marcos éticos y normativos que entienden la IA no como un tipo de dispositivo médico que debe someterse a evaluación empírica en términos de eficacia, seguridad y equidad.

La IA ha sido una pieza clave para impulsar el paradigma de la Medicina de Precisión (Sahu et al., 2022), al posibilitar la integración y el análisis de grandes volúmenes de datos clínicos, biológicos y contextuales mediante modelos predictivos que permiten estimar riesgos y probabilidades de respuesta a distintas intervenciones en pacientes concretos, desplazando así el foco de los enfoques "únicos para todos" a enfoques más personalizados. Aunque los enfoques basados en la precisión han constituido un elemento fundamental en la medicina durante décadas, su adopción en la psicología sigue siendo relativamente reciente, articulandose en el paradigma de la Salud Mental de Precisión (Bickman, 2020). Este enfoque aspira a superar los tratamientos basados en promedios grupales ("lo que funciona para la mayoría") para ofrecer una atención más personalizada, cuyo objetivo se resume en una máxima clínica: ofrecer la intervención correcta, con la intensidad adecuada, en el momento preciso y por el clínico más idóneo para cada individuo. Hasta ahora, la personalización dependía exclusivamente de la intuición clínica; actualmente la IA permite operacionalizar esta personalización procesando variables multimodales (p. ej., historia clínica, genética, contexto social y fenotipado digital) para informar esas decisiones clínicas de forma sistemática.

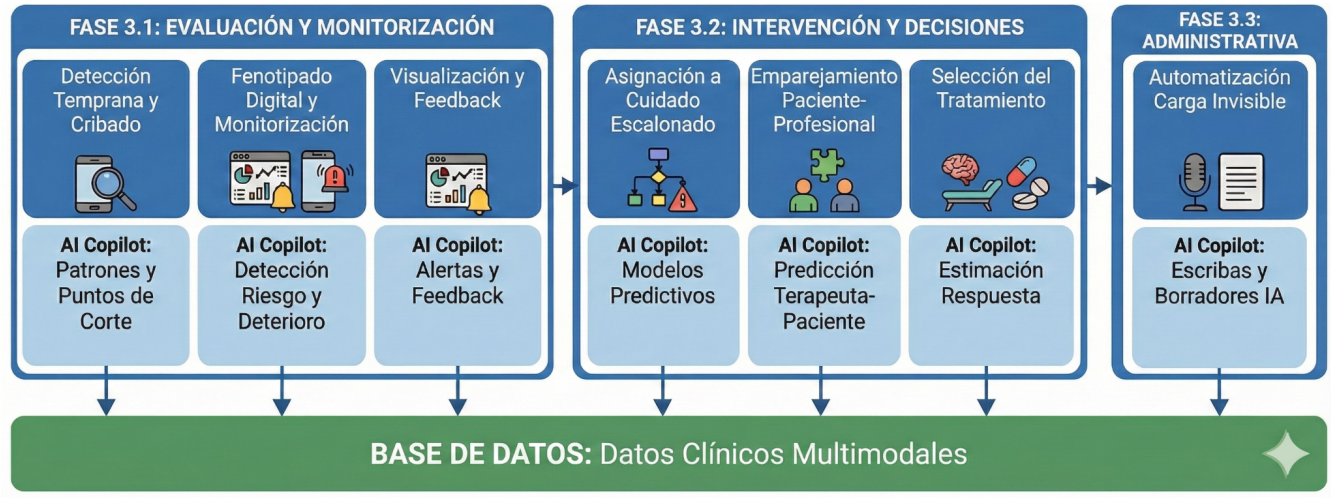
La IA en el Maletín de Herramientas Psicológicas

Si aceptamos el paradigma de Salud Mental de Precisión, la tecnología deja de ser una amenaza para convertirse en un aliado que transforma las tareas y fases críticas de la atención psicológica (figura 2):

Fase de Evaluación y Monitorización

Las herramientas de IA promueven un Cuidado Basado en la Medición (*Measurement-Based Care;* Lutz et al., 2024), que consiste en la evaluación sistemática y rutinaria de resultados y procesos clínicamente relevantes para el caso, convirtiéndose en los cimientos de la toma de decisiones basada en datos. La IA muestra

Figura 2
El Flujo de Trabajo del Psicólogo Aumentado con el Copiloto IA



un buen rendimiento en tres elementos clave de la evaluación y monitorización (Leaning et al., 2024; Spittal et al., 2025): (1) Detección temprana y cribado mediante datos rutinarios: los modelos de IA entrenados con datos de la historia clínica, cuestionarios, demográficas… son capaces de distinguir a personas con trastornos mentales de controles sanos con una gran precisión, a veces superando la sensibilidad y especificidad de los puntos de corte tradicionales; (2) Fenotipado digital y monitorización pasiva: mediante análisis pasivo de datos de teléfonos y *wearables* (p. ej., uso del móvil, movilidad, sueño, frecuencia cardíaca…), así como evaluaciones ecológicas momentáneas y muestreo de experiencias, el sistema actúa como un radar que detecta señales tempranas de deterioro, recaídas o riesgo de abandono; (3) Visualización de resultados y feedback del progreso: a partir de los datos recopilados, los modelos predictivos generan señales de alerta y pronósticos individualizados que se devuelven al clínico y al paciente en forma de *feedback* comprensible y accionable, facilitando la toma de decisiones compartida y la adaptación oportuna del plan terapéutico.

Fase de Intervención y Toma de Decisiones

Los datos recopilados durante los procesos de evaluación son procesados por modelos predictivos que transforman las mediciones en recomendaciones clínicas personalizadas, promoviendo la Toma de Decisiones Basada en Datos (*Data-Informed Decision Making;* Lutz et al., 2024). El impacto de la IA en la toma de decisiones ocurre incluso antes de que el paciente entre en consulta. Tradicionalmente, la asignación de pacientes a un determinado servicio y un determinado profesional se basa en la ubicación geográfica y la disponibilidad horaria. Sin embargo, las investigaciones sobre los Modelos de Cuidado Escalonado (*Stepped Care Model;* Scodari et al., 2023) muestran que la forma en que se realiza esta asignación inicial tiene un impacto directo en la eficiencia del sistema y en los resultados clínicos, y que las reglas explícitas basadas en la gravedad, el riesgo, las necesidades específicas y el perfil esperado de respuesta superan a los criterios organizativos tradicionales. De la misma forma, la investigación sobre el "efecto del terapeuta" demuestra que los profesionales varían significativamente en su eficacia dependiendo del tipo de paciente o patología (Johns et al., 2019), por lo que se recomienda un emparejamiento paciente-profesional basado en datos para predecir qué terapeuta tiene mayor probabilidad de establecer una alianza sólida y lograr mejores resultados con un perfil de paciente específico (Constantino et al., 2021).

Una vez seleccionado el servicio y profesional más adecuado para el caso, aparece una nueva pregunta que requiere respuesta: "¿Qué tratamiento es el más adecuado para esta persona concreta?". Esto se conoce como Selección del Tratamiento (*Treatment Selection;* Cohen and DeRubeis, 2018) y consiste en estimar, a partir de las características clínicas, personales y contextuales del paciente, qué intervención (o combinación de intervenciones) tiene mayor probabilidad de producir beneficio y menor riesgo de no respuesta o efectos adversos, con el fin de guiar la elección inicial del tratamiento y su ajuste posterior si fuera necesario. En este sentido, los modelos de IA son capaces de estimar de forma moderada quién responderá y quién no a un psicofármaco o psicoterapia, sobre todo cuando combinan información de diferentes fuentes multimodales (Lutz et al., 2025).

Fase Administrativa: La Reducción de la Carga Invisible

Quizás el beneficio más inmediato para el psicólogo saturado sea la automatización de la denominada "carga invisible". La documentación y burocracia consume un porcentaje significativo de la jornada laboral, contribuyendo al agotamiento profesional. Las nuevas herramientas de IA generativa actúan como super-secretarios: pueden transcribir sesiones (bajo estricto consentimiento y privacidad), extraer los puntos clave y redactar borradores de informes o notas de evolución en segundos. Datos preliminares indican que estos asistentes pueden reducir el tiempo administrativo entre un 40% y un 60%, devolviendo al psicólogo horas semanales que se pueden reinvertir en la atención directa, la formación continua o el autocuidado (Olson et al., 2025).

La Psicolog-IA Ante un Punto de Inflexión

La psicología se encuentra ante una decisión crítica que definirá el futuro de la profesión. Una vez demostrado que la tecnología puede procesar lenguaje y detectar patrones clínicos, la pregunta deja de ser si la IA puede ser de utilidad en salud mental, para decidir qué rol queremos que juegue dentro de la ecuación (Roca, 2025). Actualmente, observamos una tensión entre dos paradigmas de implementación contrapuestos:

El Modelo de IA Sustitutiva (Terapia Basada en IA)

Este enfoque, impulsado a menudo por una lógica de escalabilidad masiva, propone el uso de *chatbots* o aplicaciones autónomas que interactúan directamente con el usuario sin la intermediación de un profesional. Bajo la promesa de democratizar el acceso a servicios de salud mental (atención 24/7 a coste marginal cero), este modelo corre el riesgo de crear un "ecosistema paralelo" de salud mental. Aunque estas herramientas pueden ser útiles para la psicoeducación o el entrenamiento en habilidades simples (Boucher et al., 2021), plantean riesgos severos si se posicionan como terapia real: desde la dificultad para gestionar crisis y casos graves, hasta la erosión del vínculo terapéutico. El riesgo subyacente es la mercantilización automatizada del cuidado, un escenario donde la pericia clínica se ve desplazada por algoritmos que, si bien logran mimetizar la empatía, carecen fundamentalmente de juicio ético y de capacidad para interpretar la complejidad contextual del paciente (Le Glaz et al., 2021).

El Modelo de Copiloto Terapéutico (Terapia Asistida por IA)

En este modelo, la IA no se diseña para reemplazar al psicólogo, sino para "aumentar" sus capacidades ("Inteligencia aumentada"). La tecnología se integra en el flujo de trabajo clínico para encargarse de tareas computacionales (p. ej., procesar datos, detectar riesgos, organizar información…), permitiendo al profesional centrarse en las tareas relacionales e interpretativas. La analogía más clara es la del sistema de navegación GPS en la conducción. El GPS (la IA) procesa mapas y tráfico en tiempo real para sugerir la mejor ruta, pero es el conductor (el psicólogo) quien maneja el volante, decide si sigue la sugerencia y asume la responsabilidad del viaje. Bajo este prisma, la IA se convierte en una herramienta de verificación y soporte: no decide qué hacer con el paciente, pero ayuda al profesional a ver con más claridad las opciones disponibles, reducir los sesgos humanos contrastando tus decisiones con recomendaciones basadas en datos, mejorando de esta forma la eficacia y eficiencia de las intervenciones.

Ejemplo de Implementación: *Psypilot* Como Copiloto Terapéutico

Para materializar el modelo de terapia asistida por IA, en este apartado analizamos el funcionamiento de *Psypilot* como ejemplo de implementación real de un copiloto terapéutico para la psicología. *Psypilot* nace con el objetivo de convertirse en un sistema de apoyo a las decisiones clínicas en salud mental, utilizando modelos de IA para mejorar la precisión y eficiencia de los procesos de evaluación, monitorización, toma de decisiones, planificación del tratamiento y documentación, liberando carga de trabajo del profesional para que pueda centrarse en el núcleo relacional y ético de la terapia (ver figura 3).

Desde una perspectiva funcional, la herramienta operacionaliza la Salud Mental de Precisión. En lugar de depender únicamente de la entrevista clínica, el sistema despliega una evaluación de cribado validada psicométricamente que captura no solo sintomatología psicológica, sino también procesos transdiagnósticos cruciales

Figura 3
Interfaz de Psypilot Mostrando el Cuadro de Mandos de un Caso Clínico

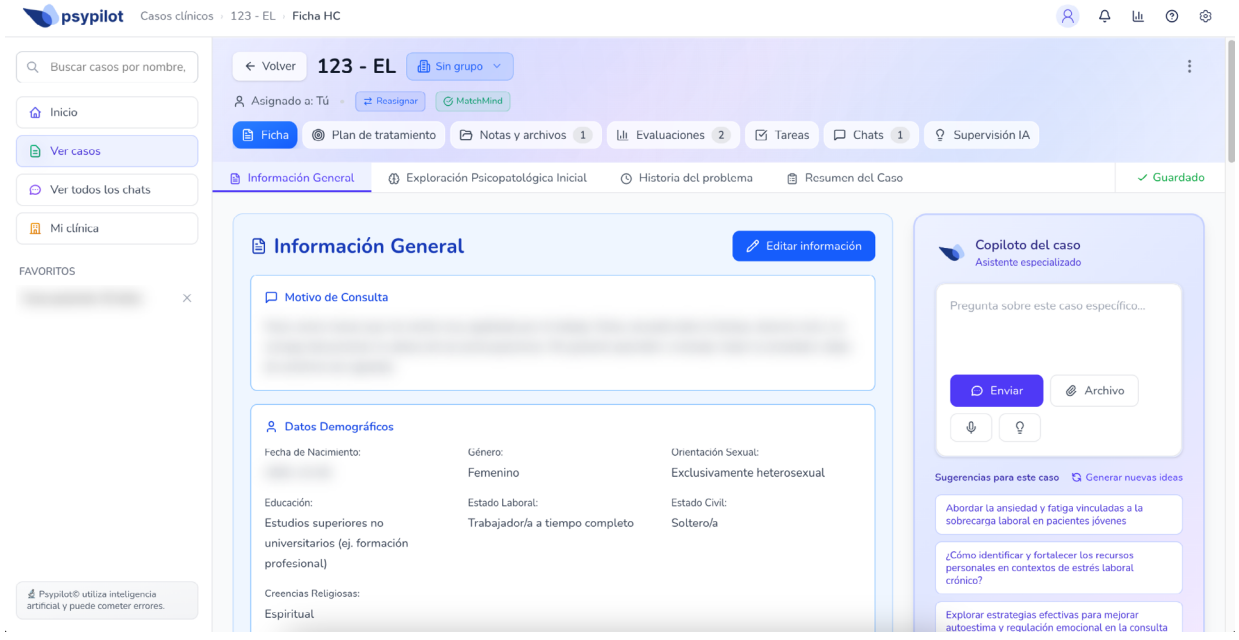
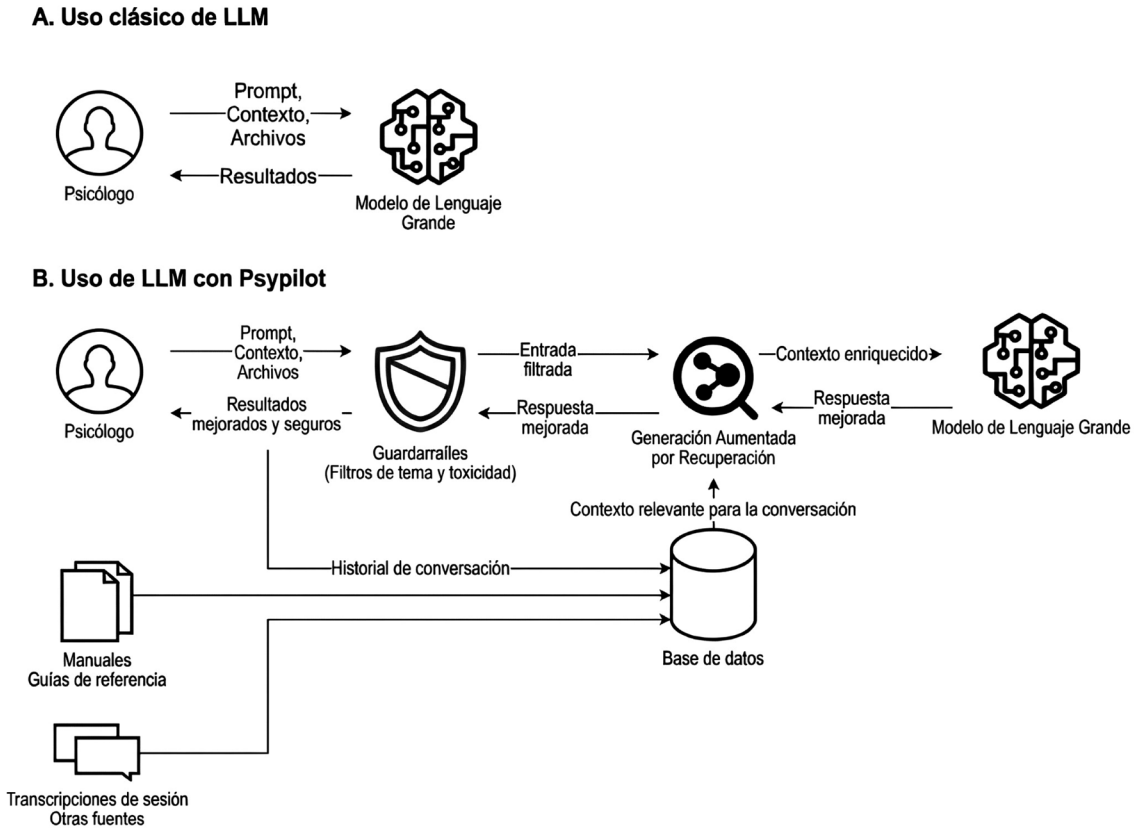


Figura 4

Comparativa de Arquitecturas. A) Uso Clásico de un LLM (Riesgo de Alucinación). B) Enfoque de Psypilot con Guardrails *de Seguridad y RAG, Enriqueciendo el Contexto con Datos Clínicos Curados para Garantizar la Veracidad*



para diversos trastornos, estilo de vida, salud general y variables contextuales. Esta información, junto con otra que puede aportar el terapeuta (p. ej., notas de sesión, ejercicios del paciente, informes de otros profesionales…), alimenta los algoritmos de personalización, que sugieren itinerarios de intervención y generan paneles visuales (*dashboards*) para monitorizar el cambio clínico durante el proceso terapéutico, facilitando una verdadera terapia basada en la medición y una toma de decisiones basada en datos.

Lo que distingue a un copiloto como *Psypilot* de herramientas generalistas como ChatGPT o Gemini es la arquitectura de seguridad. El uso de modelos de lenguaje en sanidad presenta dos riesgos críticos: la toxicidad en las respuestas y las "alucinaciones" (invención de datos). Para mitigar esto, *Psypilot* emplea una estrategia dual. Por un lado, implementa guardarraíles de seguridad (guardrails) que actúan como filtros de moderación para bloquear contenidos inapropiados o fuera de ámbito. Por otro, utiliza técnicas avanzadas como la Generación aumentada por recuperación (*Retrieval Augmented Generation* o RAG). El proceso técnico es el siguiente (ver figura 4):

1. *Contextualización:* Cuando el psicólogo realiza una consulta, el sistema no deja que el modelo "improvise" basándose en su entrenamiento general.
2. *Búsqueda vectorial:* El sistema convierte la consulta en representaciones matemáticas (embeddings) y busca la información semánticamente más relevante dentro de una base de datos vectorial cerrada y curada. Esta base de datos contiene información del caso en cuestión (p. ej., notas de sesión, datos de evaluaciones previas, tareas para casa…), así como documentos científicos de psicología basados en la evidencia científica (p. ej, guías de práctica clínica, manuales diagnósticos, protocolos de intervención), a diferencia de modelos generalistas como ChatGPT o Gemini en donde se mezcla información divulgativa y científica que está en internet sin que nadie haya realizado un filtrado previo de su validez y fiabilidad.
3. *Respuesta anclada:* El modelo genera la respuesta utilizando solo esa información verificada para mejorar la precisión de la respuesta, reducir el riesgo de alucinaciones, y asegura que las sugerencias estén alineadas con la evidencia científica y la historia real del paciente.
4. *Gobernanza de datos:* a nivel de gobernanza de datos, la infraestructura se despliega en regiones nube que garantizan el cumplimiento estricto del Reglamento General de Protección de Datos (RGPD) europeo, asegurando la soberanía del dato clínico.

Discusión: Desafíos Éticos y Regulatorios

La incorporación de sistemas de IA en psicología no puede interpretarse como una simple mejora tecnológica, sino como un cambio estructural en la forma de organizar el trabajo asistencial y sostener la calidad clínica en un contexto de creciente demanda

y recursos limitados en salud mental. En este marco, la cuestión central no es si la IA debe adoptarse, sino cómo vamos a implementarla: como un sustituto de la labor humana o como un copiloto que asista al profesional en las decisiones clínicas a lo largo del proceso terapéutico.

A diferencia de otros entornos regulatorios, el Reglamento Europeo de IA *(EU AI Act)* sitúa explícitamente los sistemas usados para triaje, diagnóstico o evaluación de riesgos en la categoría de alto riesgo (European Parliament, 2024). Este encuadre es especialmente relevante para la psicología, porque desplaza estas herramientas fuera del ámbito de "bienestar digital" y las aproxima a la lógica de producto sanitario: exige calidad de datos, trazabilidad, transparencia, gobernanza, gestión de riesgos y supervisión humana. Para el profesional, esto redefine el estándar mínimo de buena práctica: no basta con que la herramienta funcione, debe ser auditable, explicable en términos operativos y usada bajo responsabilidad profesional. Desde esta perspectiva, el caso analizado de *Psypilot* apoya la idea de que la IA puede contribuir a una salud mental de precisión si se integra en infraestructuras diseñadas para reducir fallos previsibles y proteger información sensible. En concreto, el uso de arquitecturas seguras como RAG permite acotar el modelo a fuentes controladas, disminuyendo el riesgo de alucinaciones y reforzando la seguridad clínica y la privacidad. El beneficio es doble: (a) automatización de tareas de bajo valor añadido (registro, documentación, cribado administrativo) y (b) apoyo a las decisiones clínicas en base a datos, pero preservando la centralidad del juicio profesional en todo momento.

Un hallazgo conceptual clave es que el principal peligro no reside únicamente en el error algorítmico, sino en la relación profesional-herramienta. La evidencia sobre sesgo de automatización muestra que, ante recomendaciones con apariencia de precisión (*scores,* probabilidades, riesgo estimado), los humanos tienden a sobreponderar el *output* del sistema y a infravalorar su propio juicio clínico (Shatte et al., 2019). En salud mental, este fenómeno puede ser especialmente dañino: la clínica psicológica se construye con información contextual, narrativa y relacional, difícilmente reducible a indicadores numéricos. Por ello, el modelo de IA como copiloto implementado en herramientas como *Psypilot* no es solo una preferencia práctica, sino una exigencia ética y regulatoria: el diseño debe forzar un esquema *"human-in-the-loop",* donde la IA entregue hipótesis y no dictámenes ("podría estar en riesgo de…", "patrón compatible con…"), y donde el profesional debe confirmar, rechazar o matizar activamente la sugerencia en función de la entrevista, la observación y el conocimiento del caso. Esta asimetría es crucial: la responsabilidad legal y deontológica permanece en el profesional, de modo que la interfaz, los flujos de trabajo y los protocolos deben evitar que la IA se convierta en una autoridad implícita.

La discusión sobre rendimiento no puede limitarse a métricas globales (*accuracy,* AUC) porque la IA aprende de datos que arrastran desigualdades estructurales. Si los conjuntos de entrenamiento sobrerrepresentan poblaciones occidentales, urbanas o de mayor nivel socioeconómico, el modelo puede fallar precisamente en quienes más riesgo tienen de una atención deficitaria: minorías culturales, zonas rurales o contextos con expresiones idiosincráticas del malestar (Le Glaz et al., 2021). En psicología, esto se traduce en una amenaza directa a la validez ecológica: un sistema entrenado con notas clínicas de un hospital privado urbano podría interpretar de forma incompleta (o errónea) el significado clínico de ciertas narrativas, estilos comunicativos o estresores. En consecuencia, la implementación responsable requiere una auditoría continua de sesgos y una evaluación por subgrupos (sexo/género, edad, origen cultural, nivel socioeconómico, ruralidad). La equidad debe tratarse como un criterio de calidad clínica, no como un añadido opcional, ya que optimizar el rendimiento sacrificando a grupos vulnerables sería incompatible con una práctica psicológica ética (Putica et al., 2025).

El uso de modelos complejos, especialmente en deep learning, introduce el problema de la caja negra: la dificultad (o imposibilidad) de reconstruir con claridad la cadena causal que conecta datos y predicciones. Esto genera un dilema para el consentimiento informado: si el profesional no puede explicar por qué el sistema sugiere un determinado curso de acción, ¿qué calidad tiene el consentimiento del paciente respecto al uso de herramientas de IA y su papel en la decisión clínica? (Gerke et al., 2020). La discusión aquí no es solo técnica, sino terapéutica: la confianza se construye con comprensión, y la opacidad puede erosionar el encuadre si el paciente percibe que una máquina decide. El marco regulatorio empuja hacia obligaciones de información y transparencia: el paciente debe saber cuándo se utiliza IA, para qué y con qué límites. Lejos de ocultarlo, integrar la herramienta en el encuadre puede reforzar la alianza si se presenta con honestidad, como apoyo auxiliar, no como sustituto del profesional. Además, dado el carácter sensible de los datos psicológicos, la protección bajo RGPD debe operacionalizarse en decisiones concretas: minimización de datos, control de accesos, trazabilidad, y prohibición de usos secundarios no autorizados (por ejemplo, venta de datos a terceros).

Una consecuencia transversal del uso de la IA en psicología es el desafío formativo. La discusión converge en que la adopción segura de IA exige alfabetización algorítmica para interpretar outputs (incertidumbre, límites, sesgos), supervisar su uso y contextualizar recomendaciones en la singularidad de cada paciente. Esta competencia no equivale a programar, sino a dominar una nueva herramienta clínica con criterio: cuándo confiar, cuándo dudar, cómo verificar, cómo documentar y cómo comunicar al paciente el papel de la IA en el proceso terapéutico. En España, esto sugiere integrar dicha alfabetización tanto en el grado como en la formación sanitaria especializada, y acompañarla de guías institucionales de implementación y supervisión.

Contribución a la Autoría

Pablo Roca redactó el manuscrito.

Todos los demás autores contribuyeron a la conceptualización del proyecto y también a la revisión, edición y formato final del artículo.

Financiación

Este estudio está financiado parcialmente por el proyecto PID2024-156740OA-I00 financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

Conflicto de Intereses

Los autores colaboran con una empresa que desarrolla soluciones tecnológicas en psicología, aunque afirman que el trabajo se realizó con plena independencia científica.

Referencias

Bickman, Leonard (2020). Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision mental health. *Administration and Policy in Mental Health and Mental Health Services Research, 47*(5), 795-843. https://doi.org/10.1007/s10488-020-01065-8

Boucher, Eliane M.; Harake, Nicole R.; Ward, Haley E.; Stoeckl, Sarah E.; Vargas, Junielly; Minkel, Jared; Parks, Acacia C. y Zilca, Ran (2021). Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Review of Medical Devices, 18,* 37-49. https://doi.org/10.1080/17434440.2021.2013200

Cohen, Zachary D. y DeRubeis, Robert J. (2018). Treatment selection in depression. *Annual Review Clinical Psychology, 14,* 209-236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Constantino, Michael J.; Boswell, James F.; Coyne, Alice E.; Swales, Thomas P. y Kraus, David R. (2021). Effect of matching therapists to patients vs assignment as usual on adult psychotherapy outcomes: A randomized clinical trial. *JAMA Psychiatry, 78*(9), 984-993. https://doi.org/10.1001/jamapsychiatry.2021.1221

Dwyer, Dominic B.; Falkai, Peter y Koutsouleris, Nikolaos. (2018). Machine Learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*(1), 91-118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Esteva, Andre; Kuprel, Brett; Novoa, Roberto A.; Ko, Justin; Swetter, Susan M.; Blau, Helen M. y Thrun, Sebastian (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118. https://doi.org/10.1038/nature21056

European Parliament (2024). *Artificial Intelligence Act: (2024)0138.* European Union. https://artificialintelligenceact.eu/wp-content/uploads/2024/04/TA-9-2024-0138_EN.pdf

Gerke, Sara; Minssen, Timo y Cohen, Glenm (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. En: Adam Bohr y Kaveh Memarzadeh (Eds.), *Artificial Intelligence in Healthcare* (pp. 295-336). Academic Press. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Johns, Robert G.; Barkham, Michael; Kellett, Stephen y Saxon, David (2019). A systematic review of therapist effects: A critical narrative update and refinement to Baldwin and Imel's (2013) review. *Clinical Psychology Review, 67,* 78-93 . https://doi.org/10.1016/j.cpr.2018.08.004

Lawrence, Hanna R.; Schneider, Renee. A.; Rubin, Susan B.; Matarić, Maja J.; McDuff, Daniel J. y Jones Bell, Megan (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health, 11,* e59479. https://doi.org/10.2196/59479

Le Glaz, Aziliz; Haralambous, Yanis; Kim-Dufor, Deok-Hee; Lenca, Philippe; Billot, Romain; Ryan, Taylor C.; Marsh, Jonathan; DeVylder, Jordan; Walter, Michel; Berrouiguet, Sofian y Lemey, Christophe (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research, 23*(5), e15708. https://doi.org/10.2196/15708

Leaning, Imogen E.; Ikani, Nessa; Savage, Hanna S.; Leow, Alex; Beckmann, Christian; Ruhé, Henricus. G. y Marquand, Andre F. (2024). From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. N*euroscience & Biobehavioral Reviews, 158,* 105541. https://doi.org/10.1016/j.neubiorev.2024.105541

Lee, Ellen E.; Torous, John; De Choudhury, Munmun; Depp, Colin A.; Graham, Sarah A.; Kim, Ho-Cheol; Paulus, Martin P.; Krystal, John H. y Jeste, Dilip V. (2021). Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6*(9), 856-864. https://doi.org/10.1016/j.bpsc.2021.02.001

Lutz, Wolfgang; Vehlen, Antonia y Schwartz, Brian (2024). Data-informed psychological therapy, measurement-based care, and precision mental health. *Journal of Consulting and Clinical Psychology, 92*(10), 671-673. https://doi.org/10.1037/ccp0000904

Lutz, Wolfgang; Schwartz, Brian; Vehlen, Antonia; Eberhardt, Steffen T. y Delgadillo, Jaime (2025). Advances in personalization of psychological interventions. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA), 24*(3), 343-345. https://doi.org/10.1002/wps.21342

McKinney, Scott M.; Sieniek, Marcin; Godbole, Varum; Godwin, Jonathan; Antropova, Natasha; Ashrafian, Hutan; Back, Trevor; Chesus, Mary; Corrado, Greg S.; Darzi, Ara; Etemadi, Mozziyar; Garcia-Vicente, Florencia; Gilbert, Fiona J.; Halling-Brown, Mark; Hassabis, Demis; Jansen, Sunny; Karthikesalingam, Alan; Kelly, Christopher. J.; King, Dominic; Ledsam, Joseph R.; Melnick, David; Mostofi, Hormuz; Peng, Lily; Reicher, Joshua J.; Romera-Paredes, Bernardino; Sidebottom, Richard; Suleyman, Mustafa; Tse, Daniel; Young, Kenneth C.; De Fauw, Jeffrey y Shetty, Shravya (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89-94. https://doi.org/10.1038/s41586-019-1799-6

Olson, Kristine D.; Meeker, Daniela; Troup, Matt; Barker, Timothy. D.; Nguyen, Vinh H.; Manders, Jennifer B.; Stults, Cheryl D.; Jones, Veena G.; Shah, Sachin D.; Shah, Tina y Schwamm, Lee H. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open, 8*(10), e2534976. https://doi.org/10.1001/jamanetworkopen.2025.34976

Putica, Andrea; Khanna, Rahul; Bosl, William; Saraf, Sudeep y Edgcomb, Juliet (2025). Ethical decision-making for AI in mental health: The Integrated Ethical Approach for Computational Psychiatry (IEACP) framework. *Psychological Medicine, 55,* e213. https://doi.org/10.1017/S0033291725101311

Roca, Pablo (2025). ¿Puede una mente artificial sanar una mente natural? Aplicaciones de la inteligencia artificial en psicología. En José M. Ortiz-Ibarz y Jaime Benguría-Aguirreche (Coords.), *Un nuevo conocimiento transversal: la inteligencia artificial aplicada* (pp. 145-165). Tirant lo Blanch.

Russell, Stuart y Norvig, Peter (2020). *Artificial Intelligence: A Modern Approach* (4th Ed.). Pearson.

Sahu, Mehar; Gupta, Rohan; Ambasta, Rashmi. K. y Kumar, Pravir (2022). Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. *Progress in Molecular Biology and Translational Science, 190,* 57-100. Elsevier. https://doi.org/10.1016/bs.pmbts.2022.03.002

Schaffrath, Jana; Weinmann-Lutz, Birgit y Lutz, Wolfgang (2022). The Trier Treatment Navigator (TTN) in action: Clinical case study on data-informed psychological therapy. *Journal of Clinical Psychology, 78*(10), 2016-2028. https://doi.org/10.1002/jclp.23362

Scodari, Bruno T.; Chacko, Sarah; Matsumura, Rina y Jacobson, Nicholas C. (2023). Using machine learning to forecast symptom changes among subclinical depression patients receiving stepped care or usual care. *Journal of Affective Disorders, 340,* 213-220. https://doi.org/10.1016/j.jad.2023.08.004

Shatte, Adrian B.R.; Hutchinson, Delyse M. y Teague, Samantha J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*(09), 1426-1448. https://doi.org/10.1017/S0033291719000151

Spittal, Matthew. J.; Guo, Xianglin Aneta; Kang, Laurant; Kirtley, Olivia J.; Clapperton, Angela; Hawton, Keith; Kapur, Nav; Pirkis, Jane y Carter, Greg (2025). Machine learning algorithms and their predictive accuracy for suicide and self-harm: Systematic review and meta-analysis. *PLOS Medicine, 22*(9), e1004581. https://doi.org/10.1371/journal.pmed.1004581