Intelética 3(5) (2026), 1-12
Revista de Inteligencia Artificial, Ética y Sociedad

INTELETICA

https://inteletica.iberamia.org/

La Inteligencia Artificial Generativa en los trastornos de personalidad: potencialidades, riesgos clínicos, epistemológicos y éticos

José Alberto Sotelo Martín

jose.sotelo@unir.net

Universidad Internacional de La Rioja
ORCID: https://orcid.org/0000-0002-5400-6788

Resumen Esta revisión narrativa tuvo por objetivo examinar, desde una mirada crítica e interdisciplinar, las oportunidades clínicas y los riesgos asociados al uso de inteligencia artificial generativa (IAG) en la psicoterapia de los trastornos de personalidad, considerando su aplicabilidad ética y epistemológicamente fundamentada en contextos clínicos complejos. La metodología consistió en una búsqueda bibliográfica exhaustiva en bases de datos como PubMed, Scopus y Web of Science, incluyendo artículos publicados entre 2020 y 2025. De un total de 128 estudios, se seleccionaron 26 según criterios de calidad, actualidad y relevancia temática, siguiendo directrices reconocidas para revisiones narrativas. Entre los resultados más destacados se encuentran las aplicaciones diagnósticas avanzadas de los modelos de lenguaje de gran escala (LLM), así como su uso en entornos virtuales de entrenamiento emocional. Estos avances permiten detectar patrones psicopatológicos sutiles y personalizar intervenciones. Sin embargo, también emergen riesgos significativos: reducción de la complejidad subjetiva, sesgos algorítmicos, pérdida de empatía en la interacción terapéutica y dificultades para garantizar privacidad y consentimiento informado. Como conclusión, se enfatiza que la IAG puede enriquecer la práctica psicoterapéutica siempre que no sustituya la centralidad del juicio clínico humano y se incorpore con marcos éticos sólidos, supervisión profesional activa y formación continua en competencias críticas y tecnológicas.

Abstract: This narrative review aimed to examine, from a critical and interdisciplinary perspective, the clinical opportunities and associated risks of using generative artificial intelligence (GAI) in the psychotherapy of personality disorders, considering its ethically and epistemologically grounded applicability in complex clinical settings. The methodology involved a comprehensive literature search across databases such as PubMed, Scopus, and Web of Science, including articles published between 2020 and 2025. Out of 128 studies initially identified, 26 were selected based on criteria of quality, currency, and thematic relevance, following established guidelines for narrative reviews. Among the most prominent findings were the advanced diagnostic applications of large language models (LLMs) and their use in virtual environments for emotional training. These developments make it possible to detect subtle psychopathological patterns and tailor interventions accordingly. However, significant risks also arise: a reduction in subjective complexity, algorithmic biases, loss of empathy in therapeutic interactions, and challenges in safeguarding privacy and informed consent. In conclusion, the review highlights that GAI can enhance psychotherapeutic practice as long as it does not replace the central role of human clinical judgement and is implemented within robust ethical frameworks, with active professional oversight and ongoing training in critical and technological competencies.

ISSN: 3020-7444

Palabras clave: Inteligencia artificial generativa; Trastornos de personalidad; Psicoterapia; Ética clínica; Epistemología.

Keywords: Generative artificial intelligence; Personality disorders; Psychotherapy; Clinical ethics; Epistemology.

1.- Introducción

En los últimos años, la inteligencia artificial (IA) ha experimentado un crecimiento exponencial dentro del ámbito clínico, especialmente en la psicoterapia. Este fenómeno no es casual, sino producto de una intersección entre avances tecnológicos, la creciente necesidad de optimizar recursos terapéuticos y la búsqueda constante de intervenciones más precisas y personalizadas en salud mental. La IA, en sus múltiples manifestaciones (algoritmos predictivos, sistemas de reconocimiento emocional, plataformas de interacción virtual, etc.), promete transformar la manera en que se comprenden, diagnostican y abordan trastornos psicológicos complejos, ofreciendo posibilidades hasta hace poco impensables en términos de anticipación, monitoreo y personalización terapéutica (Obradovich et al., 2024). La relevancia particular de explorar esta relación en los trastornos de personalidad radica en la naturaleza misma de estas condiciones. Estos trastornos, caracterizados por patrones persistentes y rígidos de comportamiento, pensamiento y regulación emocional, suponen un desafío considerable para la clínica tradicional debido a la dificultad para lograr diagnósticos precisos y tratamientos efectivos. Frecuentemente, estos trastornos se presentan con una alta variabilidad sintomática y patrones de recaída recurrentes, lo que hace que la intervención oportuna y adecuada sea crucial.

En este contexto, la IA ofrece un potencial significativo: algoritmos avanzados pueden detectar señales emocionales sutiles, identificar patrones conductuales problemáticos con precisión y, sobre todo, anticipar episodios críticos permitiendo así intervenciones proactivas. Ejemplos prácticos incluyen la identificación temprana de estados afectivos en el trastorno límite de personalidad, la detección automática de estados emocionales complejos como la grandiosidad y la vergüenza en el trastorno narcisista, y el análisis exhaustivo de patrones obsesivos en el trastorno obsesivo-compulsivo de personalidad (Sezgin & McKay, 2024). Sin embargo, la introducción de estas herramientas tecnológicas no está exenta de controversias ni riesgos considerables. Desde una perspectiva clínica crítica, se advierte sobre la posibilidad de que la IA pueda despersonalizar o "desubjetivar" el proceso terapéutico. La práctica clínica en psicoterapia, especialmente en la tradición psicoanalítica, se sostiene fundamentalmente sobre el reconocimiento y la validación de la experiencia subjetiva del paciente, considerando su historia personal, contexto sociocultural y la singularidad irreductible de su narrativa emocional. En contraste, la IA opera esencialmente desde un paradigma de objetivación y estandarización de la información, lo que puede conducir a un reduccionismo epistemológico al simplificar la complejidad emocional y social del paciente en datos numéricos o patrones algorítmicos (Siddals et al., 2024). Esta tensión entre subjetividad y objetividad tecnológica no es meramente teórica; tiene consecuencias éticas y epistemológicas directas. Por un lado, la precisión diagnóstica y predictiva de la IA podría promover tratamientos más efectivos y menos invasivos, evitando crisis emocionales severas y disminuyendo el sufrimiento prolongado. Por otro lado, la dependencia excesiva de algoritmos puede fomentar una práctica clínica mecánica, donde se desdibuje el vínculo terapéutico basado en la empatía, la intuición y la comprensión integral del paciente como sujeto histórico, no solo biológico o comportamental. Además, la posibilidad de que sesgos culturales, socioeconómicos o de género sean replicados inadvertidamente por algoritmos diseñados con datos históricos parcializados constituye otro desafío ético importante (Blease & Rodman, 2024).

Finalmente, mantener una perspectiva crítica sobre la integración de la IA en el tratamiento de trastornos de personalidad implica reconocer tanto las limitaciones como el potencial transformador de estas herramientas tecnológicas. El reto principal reside en cómo integrar eficazmente estos avances sin perder de vista que la psicoterapia es, ante todo, una práctica profundamente humana, donde la escucha activa, la comprensión empática y el respeto a la subjetividad del paciente son irremplazables. En consecuencia, resulta crucial que la adopción tecnológica vaya acompañada de una reflexión ética constante, capacitación adecuada de los terapeutas y un compromiso explícito con la transparencia y la explicabilidad algorítmica. Solo así será posible aprovechar plenamente las ventajas de la IA, minimizando sus riesgos y asegurando que el centro del proceso terapéutico siga siendo siempre el ser humano en su singularidad histórica y emocional.

2.- Metodología

2.1.- Estrategia de búsqueda y criterios de selección

Para esta revisión narrativa se utilizó una estrategia integral orientada a explorar y evaluar críticamente literatura reciente sobre el uso de inteligencia artificial generativa en la atención clínica de trastornos de personalidad. Se establecieron criterios específicos de selección para asegurar la relevancia, calidad y actualidad de las fuentes consultadas. Se incluyeron artículos publicados entre los años 2020 y 2025, priorizando aquellos con revisión por pares que abordaran explícitamente el uso, la eficacia, las limitaciones y los aspectos éticos y epistemológicos relacionados con la IA generativa en salud mental, especialmente enfocados en trastornos de personalidad. Se excluyeron estudios que no cumplieran con estándares científicos adecuados, aquellos que no abordaran directamente el contexto clínico, y también literatura no revisada por pares o carente de rigor metodológico suficiente. Este enfoque metodológico se fundamenta en las recomendaciones para revisiones narrativas descritas por Ferrari (2015), quien subraya la necesidad de claridad, justificación de criterios de inclusión y análisis crítico en revisiones no sistemáticas.

2.2.- Bases de datos y términos clave utilizados

La búsqueda bibliográfica se realizó en múltiples bases de datos científicas de referencia internacional: PubMed, PsycINFO, Web of Science, Scopus, SpringerLink, ScienceDirect y Google Scholar. Para optimizar los resultados y asegurar la exhaustividad del proceso, se empleó una estrategia basada en términos clave combinados mediante operadores booleanos. Las principales palabras clave utilizadas fueron: "Generative artificial intelligence", "Generative AI", "Large language models (LLM)", "Personality disorders", "Borderline personality disorder", "Ethical implications", "Clinical risks", "Epistemological concerns" y "Mental healthcare".

Se combinaron términos específicos para capturar estudios que abordaran tanto la aplicación directa de la IA generativa en intervenciones clínicas como aquellos que discutieran consideraciones éticas y epistemológicas relevantes. Esta combinación temática y estratégica sigue las recomendaciones metodológicas descritas por Green, et al. (2006), quienes destacan la importancia de seleccionar bases de datos adecuadas y de definir con precisión los términos de búsqueda en revisiones narrativas.

2.3.- Proceso de selección de estudios y análisis

Inicialmente, se recuperaron 128 artículos. Posteriormente, se llevó a cabo un proceso de cribado inicial, revisando títulos y resúmenes, lo cual permitió reducir el conjunto a 48 estudios potencialmente relevantes. En una segunda fase, se realizó una lectura crítica y exhaustiva de los textos completos de los artículos seleccionados, lo que redujo finalmente el número de estudios incluidos en esta revisión narrativa a 26. Este análisis crítico se centró en evaluar la metodología utilizada por cada artículo, la calidad y pertinencia de las conclusiones extraídas por sus autores, y su relevancia respecto a los objetivos específicos de la revisión. Se efectuó un proceso de extracción y clasificación de la información clave en una matriz estructurada, que facilitó la comparación entre estudios y permitió identificar patrones temáticos recurrentes, divergencias significativas y vacíos en el conocimiento actual. En línea con Arias Odón (2025), se valoró la validez científica de los artículos incluidos, diferenciando entre aportaciones teóricas, ensayos interpretativos y estudios empíricos con implicaciones clínicas.

2.4.- Descripción del enfoque narrativo

El enfoque adoptado para esta revisión fue narrativo-descriptivo, privilegiando la integración, síntesis y discusión crítica de la literatura seleccionada. A diferencia de las revisiones sistemáticas, este tipo de enfoque no se limita estrictamente a estudios cuantitativos, sino que también considera investigaciones cualitativas, comentarios críticos y ensayos teóricos que ofrecen perspectivas variadas y complementarias. Este método permite realizar una interpretación amplia y profunda del estado actual del conocimiento sobre el uso de IA generativa en trastornos de personalidad, identificando no solo potencialidades clínicas, sino también destacando de manera explícita y detallada riesgos clínicos, epistemológicos y éticos. Se busca así proporcionar una narrativa comprensiva que refleje tanto consensos emergentes como controversias actuales, con la finalidad de ofrecer recomendaciones

fundamentadas para futuras investigaciones e implementación práctica en contextos clínicos (Arias Odón, 2025; Ferrari, 2015; Green et al., 2006).

3.- IA Generativa y trastornos de personalidad: Contexto conceptual y epistemológico

3.1.- Definiciones fundamentales

La inteligencia artificial generativa (IAG) se refiere a sistemas tecnológicos avanzados capaces de crear contenido original e independiente, como textos, imágenes, sonidos y patrones de interacción, a partir de datos previamente analizados. Estos sistemas se basan predominantemente en algoritmos de aprendizaje automático profundo, especialmente redes neuronales generativas adversariales (GANs) y modelos de lenguaje de gran escala (LLM). Su capacidad radica en la generación autónoma de información que se asemeja considerablemente a producciones humanas, permitiendo aplicaciones diversas en contextos clínicos y psicoterapéuticos (Olawade et al., 2024). Los trastornos de personalidad, por su parte, constituyen un grupo complejo de condiciones psicológicas caracterizadas por patrones persistentes e inflexibles de comportamiento, pensamiento y regulación emocional, que frecuentemente generan dificultades en las relaciones interpersonales y en el funcionamiento social general. Ejemplos relevantes incluyen el trastorno límite de la personalidad, el trastorno narcisista y el trastorno obsesivo-compulsivo de personalidad, todos ellos con implicaciones clínicas significativas y desafíos para el diagnóstico y tratamiento (Lee, et al., 2024).

3.2.- Discusión sobre validez conceptual y metodológica

La aplicación de la IA generativa a trastornos de personalidad plantea consideraciones relevantes sobre validez conceptual y metodológica. Conceptualmente, la integración de la IA en contextos clínicos se sustenta en el supuesto de que los patrones comportamentales y emocionales pueden ser adecuadamente representados mediante datos y algoritmos predictivos. Sin embargo, esto implica desafíos epistemológicos importantes: la reducción del sujeto y su subjetividad a meros datos o patrones reconocibles puede empobrecer la comprensión integral del paciente, desconsiderando aspectos contextuales, históricos y experienciales esenciales para la práctica clínica eficaz (Springer & Smith, 2025).

Desde la perspectiva metodológica, la IA generativa ofrece una potencia predictiva sin precedentes gracias a la gran capacidad para procesar datos masivos. No obstante, esta fortaleza metodológica puede resultar problemática si no se acompaña de estrategias sólidas que mitiguen sesgos inherentes en los datos utilizados para el entrenamiento de los algoritmos. La fiabilidad y validez de los resultados generados por estos sistemas dependen críticamente de la calidad y representatividad de los datos, así como de la transparencia y explicabilidad del funcionamiento interno de los modelos generativos (Rahsepar Meadi et al., 2025). Además, es fundamental evaluar continuamente cómo estos sistemas influyen en las decisiones clínicas y terapéuticas, asegurando un enfoque equilibrado entre la precisión algorítmica y el criterio clínico profesional. Este punto es especialmente relevante en experiencias empíricas recientes que muestran cómo los usuarios valoran las intervenciones de IA generativa, pero también identifican límites en cuanto a la empatía percibida y la comprensión emocional, elementos esenciales en el abordaje de los trastornos de personalidad (Siddals et al., 2024).

4.- Potencialidades clínicas

4.1.- IA en evaluación diagnóstica (entrevistas asistidas por IA, análisis lingüístico)

Una de las áreas con mayor potencial de la IAG en psicología clínica es la evaluación diagnóstica. La capacidad de los modelos de lenguaje de gran escala para analizar patrones lingüísticos, semánticos y sintácticos en tiempo real permite detectar indicadores psicopatológicos sutiles que a menudo escapan a la percepción humana. Estudios recientes han demostrado que los algoritmos de IA pueden identificar, con elevada sensibilidad, marcadores lingüísticos asociados a la impulsividad, la labilidad afectiva o la grandiosidad narcisista, características centrales en muchos trastornos de personalidad (Roy et al., 2025). Las entrevistas asistidas por IA representan un avance

importante en este sentido. Algunos sistemas permiten al terapeuta recibir en tiempo real sugerencias diagnósticas o hipótesis clínicas preliminares basadas en el análisis automático del lenguaje verbal y no verbal del paciente. Herramientas como el análisis del ritmo de habla, el uso de pronombres, las pausas o la estructura narrativa están siendo validadas como predictores significativos en cuadros como el trastorno límite de personalidad (TLP) o el trastorno narcisista. Un ejemplo relevante es el sistema MAGI (Bi et al., 2025), que transforma entrevistas clínicas estructuradas en flujos computacionales dinámicos, ofreciendo lógica clínica y trazabilidad explícita. El análisis automatizado de textos clínicos, como historiales o notas de sesión mediante IA, permite una sistematización de la información, ayudando a detectar patrones recurrentes o cambios longitudinales relevantes para la evaluación clínica.

4.2.- Herramientas terapéuticas complementarias (entrenamiento virtual, coaching virtual)

Más allá del diagnóstico, la IAG ha comenzado a consolidarse como una herramienta terapéutica complementaria. Una de sus aplicaciones más prometedoras es el diseño de entornos virtuales de entrenamiento emocional. A través de simulaciones conversacionales controladas, los pacientes pueden ensayar respuestas ante situaciones que desencadenan sus patrones disfuncionales, lo que resulta especialmente útil en el tratamiento de trastornos de personalidad con dificultades en la regulación emocional. El llamado "coaching virtual" permite desarrollar competencias como la tolerancia a la frustración, el control de impulsos o la mentalización, apoyando el trabajo realizado en sesiones presenciales. Estos sistemas pueden estar programados para ofrecer refuerzos positivos, corregir distorsiones cognitivas o simplemente actuar como una figura de sostén entre sesiones. En contextos donde el acceso a la terapia es limitado por geografía o saturación de servicios, estos recursos constituyen un soporte valioso. Algunos dispositivos están integrando biomonitoreo (frecuencia cardiaca, sudoración, expresión facial) para adaptar en tiempo real el contenido terapéutico ofrecido por el asistente virtual, generando intervenciones personalizadas y dinámicas (Rahsepar Meadi et al., 2025).

4.3.- Ejemplos recientes de estudios experimentales y clínicos

Diversas investigaciones han comenzado a documentar empíricamente las aplicaciones clínicas de la IAG en el tratamiento de trastornos de personalidad. Siddals et al. (2024), en un estudio cualitativo, exploraron la experiencia de usuarios que interactuaban con chatbots generativos entrenados para ofrecer apoyo emocional. Los participantes informaron mejoras en la gestión de emociones, mayor introspección y sensación de compañía, aunque también destacaron la necesidad de un acompañamiento humano paralelo. Por su parte, el estudio de Lee et al. (2024) mostró que herramientas de análisis lingüístico automatizado fueron capaces de discriminar con elevada precisión entre pacientes con TLP y sujetos control, mediante el estudio de su discurso narrativo. Los resultados sugieren que estos sistemas pueden actuar como instrumentos de cribado complementarios de gran valor predictivo. Otros trabajos han evaluado el uso de entornos virtuales para la práctica de habilidades sociales en pacientes con trastorno evitativo o esquizotípico de personalidad. Se observaron mejoras significativas en indicadores de ansiedad social, conducta asertiva y empatía cognitiva, especialmente cuando las sesiones virtuales eran supervisadas por un terapeuta humano. Finalmente, Springer y Smith (2025) destacan el potencial de los modelos generativos para sintetizar información clínica compleja, ayudando al profesional a integrar datos de múltiples fuentes (entrevistas, autorregistros, biometría, cuestionarios), reduciendo la sobrecarga cognitiva y favoreciendo decisiones más informadas.

En definitiva, la IA generativa representa una herramienta de enorme valor en la psicología clínica de los trastornos de personalidad, especialmente cuando se utiliza como complemento y no como sustituto del juicio clínico humano. Sus aplicaciones en evaluación diagnóstica, análisis lingüístico y entrenamiento terapéutico abren nuevas posibilidades de intervención, personalización y acceso. No obstante, como se desarrollará en las secciones siguientes, su implementación requiere precaución metodológica, supervisión ética y una comprensión profunda de los límites y alcances reales de estas tecnologías emergentes.

5.- Riesgos clínicos

5.1.- Problemas diagnósticos automatizados: errores y reducción de la complejidad

Uno de los principales desafíos clínicos derivados del uso de inteligencia artificial generativa en contextos de evaluación psicológica es la posibilidad de una reducción excesiva de la complejidad del sujeto. Los sistemas automatizados operan mediante el reconocimiento de patrones estadísticos y correlaciones observables en los datos, lo que implica una lógica que privilegia la regularidad sobre la singularidad. Esto puede derivar en diagnósticos erróneos o imprecisos, especialmente en casos que no encajan bien en categorías clínicas rígidas o que presentan presentaciones atípicas, como ocurre frecuentemente en los trastornos de personalidad (Grabb et al., 2024).

El procesamiento algorítmico puede pasar por alto aspectos cruciales como la historia vital, los eventos traumáticos, la función simbólica del síntoma o las condiciones socioculturales del paciente. Esta reducción de la complejidad subjetiva puede conducir a decisiones clínicas basadas en indicadores fragmentarios, descontextualizados o cuantificados de manera rígida. Aunque los sistemas basados en LLM pueden identificar señales lingüísticas asociadas a psicopatología, no necesariamente comprenden el significado existencial, emocional o histórico de lo expresado (Krook, 2025). A esto se suma el riesgo de sesgos algorítmicos: si los modelos han sido entrenados sobre bases de datos clínicas históricas sesgadas por género, raza o clase, es posible que reproduzcan inequidades ya presentes en el sistema sanitario. Tal como advierten Qiu et al. (2025), incluso ligeros desbalances en los datos de entrenamiento pueden afectar significativamente la precisión diagnóstica en poblaciones clínicas específicas, con consecuencias éticas y prácticas.

5.2.- Riesgos en la interacción terapéutica: malinterpretación emocional, despersonalización

Otro ámbito crítico es el uso de IA generativa en contextos de interacción clínica. Aunque se han desarrollado chatbots y asistentes virtuales con capacidad de mantener una conversación coherente, su competencia emocional sigue siendo limitada. La IAG carece de conciencia afectiva, historia personal o capacidad de resonancia empática, elementos fundamentales en el vínculo terapéutico (Chandra et al., 2024). En pacientes con trastornos de personalidad, donde la relación terapéutica es a menudo el núcleo transformador del proceso, la malinterpretación emocional o la respuesta desintonizada pueden tener consecuencias clínicas negativas. Por ejemplo, en pacientes con trastorno límite de personalidad, una interpretación errónea del tono emocional puede desencadenar sentimientos de abandono, invalidación o rechazo. Además, el uso sistemático de IA como interlocutor terapéutico puede contribuir a un fenómeno de despersonalización: el paciente se convierte en un productor de datos, y su subjetividad es tratada como un conjunto de variables clasificables. Como señala Springer y Smith (2025), este desplazamiento hacia un enfoque tecnocrático de la salud mental puede debilitar la dimensión relacional y narrativa de la psicoterapia.

5.3.- Evidencia empírica sobre dificultades reales

Siddals et al. (2024) documentaron en su estudio cualitativo cómo algunos usuarios de chatbots generativos para salud mental experimentaban una ayuda limitada cuando sus dificultades implicaban emociones complejas o contextos interpersonales sutiles. Aunque muchos valoraron el acceso inmediato y la disponibilidad, también reportaron sensaciones de frialdad, respuestas repetitivas o fallos en la comprensión emocional. En investigaciones clínicas más amplias, como las realizadas por Rahsepar Meadi et al. (2025), se ha mostrado que la falta de transparencia de los modelos de IA, especialmente en sistemas de caja negra, impide que los clínicos comprendan cómo se llega a determinadas recomendaciones diagnósticas o terapéuticas. Esto no solo genera desconfianza, sino que limita la capacidad del profesional para ejercer un juicio clínico informado.

Por último, algunos estudios han señalado que el uso de IA en contextos de urgencia o crisis puede ser insuficiente o incluso contraproducente. Las respuestas automatizadas, aunque técnicamente correctas, pueden carecer de la contención emocional necesaria para gestionar momentos de intensa vulnerabilidad. En esos casos, el modelo puede fallar no solo en lo que dice, sino en lo que omite. En conjunto, los riesgos clínicos asociados al uso de IA generativa en trastornos de personalidad no deben ser subestimados. Aunque estas herramientas pueden complementar valiosamente ciertas tareas clínicas, su aplicación directa en el diagnóstico y la interacción terapéutica conlleva

desafíos que afectan la precisión, la seguridad emocional y la ética del acto clínico. Una supervisión humana constante, una formación ética adecuada y una comprensión crítica de los límites de estas tecnologías son condiciones indispensables para su integración responsable en la práctica psicoterapéutica.

6.- Riesgos epistemológicos

6.1.- Opacidad y dificultad en la interpretación de modelos generativos

Una de las principales preocupaciones epistemológicas en torno al uso de inteligencia artificial generativa en el ámbito clínico radica en la opacidad inherente de muchos de sus modelos. La mayoría de los grandes modelos de lenguaje funcionan como sistemas de "caja negra", en los que no es posible rastrear con claridad los procesos que conducen a un determinado resultado o sugerencia diagnóstica. Esta falta de explicabilidad compromete gravemente la posibilidad de validar científicamente los resultados, evaluar su fiabilidad o cuestionar sus implicaciones clínicas (Ji et al., 2023).

En el campo de la psicoterapia, donde la interpretación de los síntomas, las narrativas y las interacciones está profundamente contextualizada, la ausencia de trazabilidad limita la capacidad del profesional para integrar adecuadamente las recomendaciones de la IA con su propio juicio clínico. La imposibilidad de comprender cómo un modelo ha llegado a identificar un patrón de comportamiento como patológico reduce la legitimidad epistemológica de sus conclusiones y obstaculiza la posibilidad de someterlas a crítica o debate profesional (Hua et al., 2025).

6.2.- Problemas metodológicos: validez, fiabilidad y sesgos implícitos

A nivel metodológico, los sistemas de IAG enfrentan importantes desafíos en cuanto a la validez y la fiabilidad de sus outputs. Aunque pueden ofrecer predicciones con alta precisión estadística, estas predicciones no siempre tienen un valor clínico claro o replicable. La validez de una inferencia diagnóstica, por ejemplo, no solo depende de su ajuste a patrones estadísticos, sino de su capacidad para representar adecuadamente la experiencia del paciente, su historia vital y su contexto social. Los modelos generativos, al priorizar la eficiencia computacional, corren el riesgo de sacrificar la riqueza fenomenológica de los datos clínicos (Suenghataiphorn et al., 2025).

Además, muchos de estos modelos han sido entrenados con bases de datos clínicas históricamente sesgadas. Esto introduce sesgos implícitos que tienden a reproducirse y amplificarse en los resultados. Así, se corre el riesgo de reforzar visiones patologizantes, estigmatizantes o culturalmente insensibles, comprometiendo la equidad en la práctica clínica (Daneshjou et al., 2023). La fiabilidad, por su parte, queda comprometida cuando los modelos reaccionan de manera inestable a entradas mínimamente distintas, lo cual plantea serias dudas sobre su utilidad como herramienta de apoyo en decisiones terapéuticas complejas (Chen et al., 2024).

6.3.- Limitaciones en la construcción del conocimiento clínico

El despliegue de sistemas de IAG en salud mental no solo tiene efectos operativos, sino también consecuencias profundas sobre la forma en que se construye el conocimiento clínico. Cuando se privilegia la cuantificación algorítmica como criterio de verdad, se debilita el valor de otros modos de saber fundamentales en la práctica psicoterapéutica, como la interpretación subjetiva, la comprensión narrativa o el pensamiento clínico situado (Richard, 2025). Este giro epistémico puede contribuir a una forma de medicalización tecnocrática del sufrimiento psíquico, en la que la complejidad de la experiencia humana es reducida a outputs computables. Además, existe el riesgo de que el conocimiento generado por la IA se naturalice como neutral y objetivo, invisibilizando las decisiones humanas y culturales que intervienen en su diseño, entrenamiento e interpretación. Esta fetichización de la tecnología no solo empobrece el saber clínico, sino que pone en peligro la reflexividad crítica necesaria para su evolución.

En resumen, los riesgos epistemológicos asociados al uso de IAG en psicoterapia no son secundarios ni meramente técnicos. Se trata de desafíos que interpelan directamente las bases del conocimiento clínico, su legitimación y su capacidad para hacer justicia a la complejidad del sufrimiento humano. Reconocer estos límites y discutirlos abiertamente es condición indispensable para una integración ética, rigurosa y crítica de la inteligencia artificial en los contextos terapéuticos contemporáneos.

7.- Cuestiones éticas

7.1.- Privacidad, confidencialidad y consentimiento informado

El uso de inteligencia artificial generativa en psicoterapia plantea exigencias inéditas sobre la gestión ética de la privacidad y la confidencialidad. En los sistemas tradicionales, el terapeuta asume la responsabilidad de resguardar la información del paciente dentro de un marco normativo bien establecido. Sin embargo, cuando se utilizan plataformas de IA, ya sea para evaluación, intervención o seguimiento, los datos del paciente pueden almacenarse, procesarse y transferirse a través de servidores externos, muchas veces sin control directo del profesional tratante.

El consentimiento informado se vuelve, por tanto, un acto complejo que debe superar el modelo formalista de la firma documental. Los pacientes deben entender no solo qué datos serán utilizados, sino cómo serán procesados, con qué fines, quiénes tendrán acceso a ellos y qué riesgos implica su uso. La opacidad tecnológica, sumada a la asimetría informativa entre pacientes y desarrolladores, debilita la transparencia y compromete la autonomía real del usuario para consentir de manera libre e informada (Mandal et al., 2025; Hua et al., 2025). Este aspecto resulta especialmente sensible en pacientes con trastornos de personalidad, que pueden tener dificultades en la autorregulación emocional, la desconfianza relacional o la percepción de agencia.

7.2.- Responsabilidad profesional y dilemas éticos

La delegación de funciones clínicas en sistemas de IA genera una zona difusa en cuanto a la responsabilidad profesional. Si una recomendación algorítmica conduce a una decisión errónea o a un efecto adverso, ¿quién responde éticamente?, ¿El profesional que utilizó la herramienta?, ¿El equipo que diseñó el modelo?, ¿La institución que lo adoptó? Esta fragmentación de la responsabilidad pone en cuestión el principio de responsabilidad que rige la práctica clínica (Grabb et al., 2024).

Los dilemas éticos también se amplifican cuando los algoritmos realizan inferencias que el clínico no puede verificar, pero que parecen "razonables" según patrones estadísticos. La tentación de delegar el juicio clínico en una entidad que ofrece respuestas rápidas, consistentes y avaladas por datos puede erosionar la función reflexiva del terapeuta y fomentar una práctica guiada por automatismos más que por comprensión (Ma et al., 2024). Además, existe el riesgo de que las decisiones clínicas se vean condicionadas por imperativos técnicos o económicos más que por criterios éticos. Por ejemplo, la presión institucional para implementar IA por eficiencia o reducción de costes puede llevar a aceptar márgenes de error que serían inaceptables en un contexto humano. La ética profesional requiere resistir estas presiones y reafirmar la primacía del cuidado individualizado y del juicio ético contextual (Suenghataiphorn et al., 2025).

7.3.- Autonomía y capacidad decisoria del paciente

La IA generativa, al ofrecer recomendaciones personalizadas, podría ser vista como una aliada del empoderamiento del paciente. Sin embargo, en la práctica, su uso puede afectar negativamente la autonomía si no se gestiona con criterios claros de supervisión y participación. Cuando el paciente percibe que las decisiones clínicas son dictadas por una máquina, puede sentirse desplazado del centro de la relación terapéutica. La confianza en la alianza terapéutica se debilita si el paciente duda de si el profesional está actuando por comprensión empática o por obediencia al algoritmo. Esta duda es aún más crítica en personas con trastornos de personalidad, para quienes la relación con la autoridad, la validación del self y la percepción del otro como aliado o invasor son núcleos estructurantes de la experiencia subjetiva. Por tanto, preservar la autonomía no significa solo obtener un consentimiento inicial, sino construir condiciones de diálogo permanente donde el paciente pueda cuestionar,

entender y negociar el uso de la IA en su proceso terapéutico. Esto incluye el derecho a no utilizar IA, el acceso a explicaciones comprensibles y el reconocimiento de que el conocimiento técnico nunca debe eclipsar el respeto a la singularidad y dignidad de cada persona (Asman et al., 2025).

En suma, las cuestiones éticas que emergen con la implementación de IA generativa en psicoterapia no pueden abordarse como simples ajustes normativos. Requieren una reflexión profunda sobre la dignidad, la responsabilidad y la subjetividad en contextos clínicos. Solo una ética relacional, informada y vigilante podrá acompañar el uso de estas tecnologías sin traicionar el núcleo humanista de la práctica terapéutica (Rahsepar Meadi et al., 2025; Chen et al., 2024).

8.- Discusión crítica

La integración de la inteligencia artificial generativa en el campo de la psicoterapia aplicada a los trastornos de personalidad presenta un panorama complejo, en el que coexisten oportunidades innovadoras y desafíos éticos y epistemológicos sustanciales. En un recorrido que abarca desde la evaluación diagnóstica automatizada hasta la implementación de asistentes conversacionales y entornos terapéuticos virtuales, la literatura reciente ofrece perspectivas complementarias —y en ocasiones discrepantes— sobre los alcances y límites de estas tecnologías emergentes.

Desde una mirada optimista, autores como Obradovich et al. (2024) y Lee et al. (2024) destacan el potencial de los modelos de lenguaje de gran escala para mejorar la precisión diagnóstica, personalizar la atención y optimizar procesos clínicos. En consonancia, Bi et al. (2025) demuestra cómo herramientas como MAGI permiten transformar entrevistas clínicas estructuradas en flujos computacionales con trazabilidad lógica, facilitando una exploración diagnóstica más estructurada. Siddals et al. (2024), desde un enfoque cualitativo, aportan evidencia empírica sobre experiencias positivas de usuarios con chatbots generativos, subrayando beneficios como el apoyo emocional, la introspección y la mejora en la percepción relacional, aunque advierten que estos sistemas deben complementar y no reemplazar la atención humana.

A este consenso sobre las oportunidades se suman Springer y Smith (2025), quienes sostienen que los modelos generativos pueden integrar múltiples fuentes de información clínica, reduciendo la sobrecarga del profesional. Asimismo, Sezgin e Ian McKay (2024) argumentan que la IAG permite construir intervenciones terapéuticas personalizadas, especialmente cuando se integra la retroalimentación humana en el proceso de toma de decisiones. No obstante, esta perspectiva positiva se ve matizada por voces críticas que llaman a una implementación cautelosa. Blease y Rodman (2024), desde un marco de ética biomédica, cuestionan la legitimidad del conocimiento generado por la IAG cuando no se puede auditar su funcionamiento interno. Plantean que la integración de estas herramientas en salud mental debe sustentarse en principios éticos informados por evidencia empírica, con especial atención a los daños potenciales asociados.

Rahsepar Meadi et al. (2025) profundizan en los dilemas éticos específicos del uso de IA conversacional en salud mental, subrayando que la representación de la subjetividad en los modelos generativos aún es limitada y plantea desafíos para preservar la narrativa singular del paciente. En línea similar, Chen et al. (2024) alertan sobre los riesgos de privacidad y la seguridad de datos clínicos en contextos digitales, especialmente cuando la recolección, almacenamiento y procesamiento de información se externaliza a servidores no regulados por el profesional tratante.

En el plano clínico, Chandra et al. (2024) y Qiu et al. (2025) describen cómo la falta de resonancia emocional de la IAG puede derivar en interacciones mecánicas o malinterpretaciones afectivas, afectando la alianza terapéutica. Estas preocupaciones encuentran eco en Richard, (2025), quien advierte que la sustitución del vínculo humano por interfaces algorítmicas puede erosionar los fundamentos relacionales de la psicoterapia.

Desde una perspectiva epistemológica, Ji et al. (2023) y Hua et al. (2025) coinciden en la necesidad de desarrollar modelos explicables e interpretables. Critican la lógica de “caja negra” de los LLM, ya que su opacidad limita la capacidad del clínico para someter a crítica los outputs generados, comprometiendo la trazabilidad y validez del conocimiento producido. Esta línea crítica es reforzada por Suenghataiphorn et al. (2025), quienes documentan

cómo los sesgos históricos incorporados a las bases de entrenamiento tienden a replicarse, ampliando desigualdades estructurales en la atención.

En el plano institucional, Grabb et al. (2024) plantean que la implementación de IA en salud mental requiere una arquitectura ética clara sobre quién es responsable en caso de error clínico derivado del uso de estas tecnologías. Mandal et al. (2025) subrayan que la privacidad del paciente debe situarse como eje rector en el desarrollo de modelos de IA para salud mental. A ello se suma el análisis de Ma et al. (2024), quienes afirman que no existe un código ético universal para la relación entre bots y pacientes, y que toda implementación debe considerar el contexto cultural y clínico local.

Frente a este panorama, se impone un equilibrio pragmático y ético. La IAG no debe ser desestimada, pero tampoco asumida como una panacea tecnológica. Como sostiene Asman et al. (2025), su incorporación a la práctica clínica requiere diseño responsable, integración cuidadosa y supervisión activa. Esto implica establecer marcos de gobernanza regulada, diseñar protocolos de consentimiento dinámico y garantizar la capacitación continua de los profesionales sobre el funcionamiento y los límites de estas herramientas.

La investigación futura debería explorar, entre otros aspectos, la eficacia diferencial de terapias asistidas por IA según tipo de trastorno de personalidad, el impacto emocional de la interacción humano-máquina, y el desarrollo de estándares internacionales sobre explicabilidad, seguridad y justicia algorítmica. Como proponen Daneshjou et al. (2023), no se trata solo de mejorar la tecnología, sino de mejorar el ecosistema en el que esta se inserta.

En síntesis, los artículos analizados coinciden en que la IA generativa puede enriquecer la psicoterapia de los trastornos de personalidad, siempre que no desplace la centralidad del vínculo humano ni sustituya la complejidad del juicio clínico. El reto no es solo técnico, sino profundamente ético, epistemológico y político. Integrar IA en psicoterapia será posible si, y solo si, la inteligencia artificial se convierte en una extensión crítica y no reductiva de la inteligencia clínica.

9.- Conclusiones

La presente revisión narrativa ha permitido explorar, desde una perspectiva crítica e interdisciplinar, las potencialidades y riesgos del uso de inteligencia artificial generativa (IAG) en la psicoterapia de los trastornos de personalidad. Los hallazgos indican que la IAG ofrece aplicaciones prometedoras en la evaluación diagnóstica, el análisis de patrones lingüísticos, el acompañamiento terapéutico a través de chatbots, y la personalización de intervenciones clínicas. Herramientas como los modelos de lenguaje de gran escala (LLM) han demostrado ser útiles para la detección temprana de signos psicopatológicos y la sistematización de la información clínica, contribuyendo así a mejorar la eficiencia y accesibilidad del sistema de salud mental.

Sin embargo, estos avances deben ser valorados con cautela. Los riesgos clínicos derivados de la automatización del diagnóstico, la pérdida de riqueza subjetiva en la interacción terapéutica y la falta de transparencia de los modelos generativos configuran escenarios que requieren atención inmediata. Asimismo, los riesgos epistemológicos, como la opacidad algorítmica y los sesgos en los datos de entrenamiento, comprometen la fiabilidad y validez del conocimiento clínico producido por estas tecnologías. Las preocupaciones éticas sobre privacidad, consentimiento informado, autonomía del paciente y responsabilidad profesional atraviesan todos los niveles de análisis. Integrar la IAG en la práctica psicoterapéutica exige, por tanto, el establecimiento de marcos normativos claros, una gobernanza ética robusta y la formación continua de profesionales en competencias digitales, críticas y éticas. La inteligencia artificial no debe sustituir la inteligencia clínica, sino complementarla. El centro de la intervención sigue siendo la subjetividad del paciente, su historia, su contexto y su dignidad.

En definitiva, esta revisión subraya la necesidad de avanzar hacia una inteligencia artificial humanamente significativa: una IA que se articule con el juicio clínico, respete la complejidad psíquica y contribuya a una práctica terapéutica más justa, eficaz y ética.

10. Referencias

Arias-Odón, F. (2025). El artículo de revisión narrativa: nivel de evidencia y validez científica. Revisión semi-sistemática. *e-Ciencias de la Información, 15*(1), 1–24. https://doi.org/10.15517/eci.v15i1.59584

Asman, O., Torous, J., & Tal, A. (2025). Responsible design, integration, and use of generative AI in mental health. *JMIR Mental Health, 12*, e70439. https://doi.org/10.2196/70439JMIR Salud Mental

Bi, G., Chen, Z., Liu, Z., Wang, H., Xiao, X., Xie, Y., Zhang, W., Huang, Y., Chen, Y., Peng, L., Feng, Y., & Huang, M. (2025). *MAGI: Multi-Agent Guided Interview for Psychiatric Assessment*. arXiv. https://doi.org/10.48550/arXiv.2504.18260

Blease, C., Rodman, A. (2025). Generative Artificial Intelligence in Mental Healthcare: An Ethical Evaluation. *Curr Treat Options Psych* 12, 5. https://doi.org/10.1007/s40501-024-00340-x

Chandra, M., Naik, S., Ford, D., Okoli, E., De Choudhury, M., Ershadi, M., Ramos, G., Hernandez, J., Bhattacharjee, A., & Warreth, S. (2024). From lived experience to insight: Unpacking the psychological risks of using AI conversational agents. *arXiv*. https://doi.org/10.48550/arXiv.2412.07951

Chen, D., Liu, Y., Guo, Y., & Zhang, Y. (2024). The revolution of generative artificial intelligence in psychology: The interweaving of behavior, consciousness, and ethics. *Acta Psychologica, 251*, 104593. https://doi.org/10.1016/j.actpsy.2024.104593

Daneshjou, R., Omiye, T., Lester, J., Spichak, S., Rotemberg, V., & Omiye, J. A. (2023). Large language models propagate race-based medicine. *NPJ Digital Medicine, 6*(195). https://doi.org/10.1038/s41746-023-00939-z

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing, 24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Grabb, D., Lamparth, M., & Vasan, N. (2024). Risks from language models for automated mental healthcare: Ethics and structure for implementation. *arXiv*. https://doi.org/10.48550/arXiv.2406.11852

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine, 5*(3), 101–117. https://doi.org/10.1016/S0899-3467(07)60142-6

Hua, Y., Na, H., Li, Z., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. *npj Digit. Med.* 8, 230 (2025). https://doi.org/10.1038/s41746-025-01611-4

Ji, S., Zhang, T., Yang, K., Ananiadou, S., & Cambria, E. (2023). Rethinking large language models in mental health applications. *arXiv*. https://doi.org/10.48550/arXiv.2311.11267

Krook, J. (2025). Manipulation and the AI Act: Large language model chatbots and the danger of mirrors. *arXiv*. https://doi.org/10.48550/arXiv.2503.18387

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., Paulus, M. P., Krystal, J. H., & Jeste, D. 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

Ma, L., Zhao, T., Qiu, H., & Lan, Z. (2024). No general code of ethics for all: Ethical considerations in human-bot psycho-counseling. *arXiv*. https://doi.org/10.48550/arXiv.2404.14070

Mandal, A., Chakraborty, T., & Gurevych, I. (2025). Towards privacy-aware mental health AI models: Advances, challenges, and opportunities. *arXiv*. https://doi.org/10.48550/arXiv.2502.00451

Obradovich, N., Khalsa, S. S., Khan, W., Suh, J., Perlis, R. H., Ajilore, O., & Paulus, M. P. (2024). Opportunities and Risks of Large Language Models in Psychiatry. *NPP - digital psychiatry and neuroscience*, *2*(1), 8. https://doi.org/10.1038/s44277-024-00010-z

Olawade, D. B., Wada, O, Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health, 3.* https://doi.org/10.1016/j.glmedi.2024.100099

Qiu, J., He, Y., Juan, X., Wang, Y., Liu, Y., Yao, Z., Wu, Y., Jiang, X., Yang, L., & Wang, M. (2025). EmoAgent: Assessing and safeguarding human-AI interaction for mental health safety. *arXiv*. https://doi.org/10.48550/arXiv.2504.09689

Rahsepar Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the ethical challenges of conversational AI in mental health care: Scoping review. *JMIR Mental Health, 12*, e60432. https://doi.org/10.2196/60432

Richards, D. (2024). Artificial intelligence and psychotherapy: A counterpoint. *Counselling and Psychotherapy Research, 24*(1), 1–6. https://doi.org/10.1002/capr.12758

Roy, K., Surana, H., Eswaramoorthi, D., Zi, Y., Palit, V., Garimella, R., & Sheth, A. (2025). *Large Language Models for Mental Health Diagnostic Assessments: Exploring the Potential of Large Language Models for Assisting with Mental Health Diagnostic Assessments -- The Depression and Anxiety Case*. arXiv. https://doi.org/10.48550/arXiv.2501.01305

Sezgin, E., McKay, I. (2024). Behavioral health and generative AI: a perspective on future of therapies and patient care. *npj Mental Health Reearchs,* 3, 25. https://doi.org/10.1038/s44184-024-00067-w

Siddals, S., Torous, J. & Coxon, A. (2024). It happened to be the perfect thing: experiences of generative AI chatbots for mental health. *npj Mental Health Res* 3, 48. https://doi.org/10.1038/s44184-024-00097-4

Springer, S., & Smith, J. (2025). Generative AI in mental healthcare: An ethical review. *AI & Society, 40*(1), 123–135. https://doi.org/10.1007/s40501-024-00340-x

Suenghataiphorn, T., Tribuddharat, N., Danpanichkul, P., & Kulthamrongsri, N. (2025). Bias in large language models across clinical applications: A systematic review. *arXiv*. https://doi.org/10.48550/arXiv.2504.02917