Inteligencia artificial en salud mental: seguridad del paciente, ética y derecho.

AUTORES:
Arroyo Giuliani, Sandra Susana
Psicóloga y Especialista en Dependencia Emocional y Adicciones
Torres Ponce, Mariano Enrique
Abogado y Especialista en Derecho Informático

RESUMEN

Contextualización: Este trabajo analiza críticamente el uso de inteligencia artificial en salud mental, con énfasis en sus limitaciones técnicas, éticas y regulatorias. A partir de la revisión de literatura y casos ilustrativos, se examina cómo la falta de comprensión semántica, los sesgos algorítmicos y la incapacidad para captar matices culturales limitan la posibilidad de que estas herramientas funcionen como alternativas a la psicoterapia. Se destacan riesgos específicos en poblaciones vulnerables, como adolescentes, personas con trastornos de personalidad y víctimas de trauma complejo, para quienes la relación terapéutica humana resulta insustituible. En paralelo, se identifican áreas de aplicación responsable, incluyendo la psicoeducación, el triaje clínico y el apoyo al trabajo profesional. Desde la perspectiva jurídica y deontológica, se subraya la necesidad de marcos normativos claros que garanticen transparencia, consentimiento informado y supervisión profesional. El artículo concluye que el futuro de la salud mental no debe plantearse como una oposición entre tradición y tecnología, sino como un esfuerzo de integración prudente, donde la innovación digital complemente la práctica clínica sin reemplazar su núcleo humano.

ABSTRACT

Context: This paper critically examines the use of artificial intelligence in mental health, focusing on its technical, ethical, and regulatory limitations. Drawing on literature review and illustrative case studies, it explores how the lack of semantic understanding, algorithmic bias, and inability to grasp cultural nuances restrict the potential of these tools to function as genuine alternatives to psychotherapy. Particular attention is given to vulnerable populations, such as adolescents, individuals with personality disorders, and victims of complex trauma, for whom the human therapeutic relationship remains irreplaceable. At the same time, the article identifies areas for responsible integration, including psychoeducation, clinical triage, and professional support. From legal and ethical perspectives, it emphasizes the need for clear regulatory frameworks ensuring transparency, informed consent, and professional oversight. The paper concludes that the future of mental health care should not be conceived as a conflict between tradition and technology but as a prudent integration effort, where digital innovation complements clinical practice without replacing its human core.

PALABRAS CLAVE

Inteligencia artificial, salud mental, psicología clínica, relación terapéutica, ética profesional, bioética, derecho y tecnología, regulación de IA, sesgos algorítmicos, confidencialidad de datos.

KEYWORDS

Artificial intelligence, mental health, clinical psychology, therapeutic relationship, professional ethics, bioethics, law and technology, AI regulation, algorithmic bias, data confidentiality.

RESUMEN EJECUTIVO

Background: El uso de inteligencia artificial en salud mental crece con rapidez mientras persisten límites técnicos, éticos y regulatorios. La práctica clínica muestra que el cambio terapéutico se sostiene en la relación humana y en una comprensión situada del sufrimiento. Los sistemas actuales operan con correlaciones y carecen de comprensión semántica y sensibilidad cultural, lo que restringe su valor como sustitutos del terapeuta.

Gap: La evidencia favorable es breve y de corto plazo. Abundan métricas de satisfacción y uso, faltan seguimientos longitudinales, auditorías independientes y reportes de eventos adversos. El marco jurídico y deontológico es desigual, con zonas grises en responsabilidad, consentimiento y protección de datos que permiten ofertas que imitan psicoterapia sin garantías equivalentes.

Purpose: Delimitar con precisión dónde la IA aporta valor en salud mental y dónde introduce riesgo, y proponer criterios clínicos, éticos y regulatorios para su integración responsable sin sustituir el núcleo relacional de la psicoterapia.

Methodology: Revisión crítica de literatura académica y normativa reciente, análisis conceptual de límites técnicos y sesgos, y construcción de casos clínicos hipotéticos para examinar riesgo en poblaciones vulnerables. Síntesis de buenas prácticas y de modelos de regulación proporcional por nivel de riesgo.

Results: Identificación de límites estructurales en comprensión de matices, detección de engaño y representación cultural. Riesgos acentuados en adolescentes, trastornos de personalidad y trauma complejo. Áreas de uso responsable en psicoeducación, triaje y apoyo profesional. Recomendaciones para práctica clínica con consentimiento informado continuo, rutas de escalado y medición con retroalimentación. Propuestas regulatorias sobre categorización por riesgo, transparencia operativa y supervisión profesional obligatoria.

Conclusion: La IA no reemplaza la relación terapéutica. Su aporte es complementario si opera con supervisión clínica, trazabilidad y límites claros. Se prioriza inversión en atención humana, adopción gradual con evaluación independiente y una agenda de investigación que incluya estudios longitudinales, explicabilidad útil para la clínica y validación transcultural.

ÍNDICE TEMÁTICO

A. INTRODUCCIÓN

La expansión de la inteligencia artificial en salud mental atrae por su promesa de ampliar acceso y reducir barreras de tiempo y costo. Chatbots terapéuticos, herramientas de cribado automatizado y sistemas de respuesta ante crisis se ofrecen como apoyos disponibles a cualquier hora, con inmediatez seductora para usuarios y decisores (Torous y Roberts, 2017). Sin embargo, la práctica clínica muestra que la tecnología no opera en el vacío. Allí donde hay sufrimiento, historia personal y vínculos frágiles, las soluciones técnicas interactúan con realidades humanas complejas y pueden producir efectos ambivalentes. La intersección entre lo clínico, lo ético y lo jurídico exige una lectura atenta que preserve aquello que realmente produce cambio terapéutico (Fiske, Henningsen y Buyx, 2019).

Este artículo explora en qué condiciones la inteligencia artificial agrega valor a la atención psicológica sin desplazar la relación que sostiene el proceso de cambio. El análisis parte de una premisa clara: la tecnología puede colaborar con tareas acotadas como psicoeducación, seguimiento y organización de información, pero no reemplaza la deliberación compartida ni la sintonía fina que ocurren en el encuentro terapéutico. La creciente oferta comercial que presenta a los sistemas conversacionales como alternativa de sustitución obliga a examinar sus límites y a describir con cuidado los riesgos de una automatización que desatienda la fragilidad de poblaciones vulnerables (Mohr, Burns, Schueller, Clarke y Klinkman, 2013).

La literatura acumulada ubica a la relación terapéutica en el centro del cambio clínico. Los resultados dependen en gran medida de la calidad de la alianza, del reconocimiento de matices afectivos y de la capacidad para sostener una narrativa de sentido que se construye en el tiempo. Estas dinámicas exceden la aplicación de protocolos o ejecución de algoritmos y se apoyan en habilidades interpretativas y relacionales del profesional (Lambert y Barley, 2001). Sobre ese trasfondo, preocupa la difusión de mensajes que atribuyen a los sistemas artificiales equivalencia funcional con el trabajo humano, ya que elevan expectativas y pueden inducir decisiones inadecuadas en contextos de alta vulnerabilidad.

El objetivo central es distinguir con claridad entre asistencia tecnológica y sustitución terapéutica. Se propone un marco conceptual que ordena funciones, delimita responsabilidades y orienta decisiones de uso responsable. El texto aborda riesgos

técnicos, implicancias éticas y exigencias jurídicas, con la mira puesta en proteger a las personas y en ayudar a profesionales y reguladores a tomar decisiones informadas.

En el contexto de América Latina, esta promesa tecnológica adquiere matices particulares. La región enfrenta una doble realidad: por un lado, una brecha significativa en el acceso a servicios de salud mental, con un déficit de profesionales y una alta estigmatización que limita la búsqueda de ayuda. Por otro lado, una creciente penetración digital que convive con una persistente brecha digital, donde el acceso a la tecnología no es equitativo. Es en esta encrucijada donde las soluciones de IA se presentan como una alternativa atractiva, pero también donde los riesgos de reproducir y amplificar desigualdades estructurales son más pronunciados.

A.1. ENFOQUE METODOLÓGICO

Este trabajo constituye una revisión crítica interdisciplinaria con enfoque narrativo-integrativo orientada a síntesis conceptual antes que a agregación cuantitativa de estudios. La estrategia de búsqueda combinó revisión selectiva de literatura académica publicada entre 2015 y 2024 en PubMed, PsycINFO y Scopus mediante términos como artificial intelligence, mental health, chatbot, therapeutic alliance, algorithmic bias y patient safety, con análisis de documentos normativos de organismos como la Organización Mundial de la Salud, la Comisión Europea, la UNESCO y la American Psychological Association.

Los criterios de inclusión priorizaron estudios empíricos revisados por pares sobre efectividad y seguridad de sistemas conversacionales en salud mental, meta-análisis sobre intervenciones digitales, trabajos conceptuales sobre límites técnicos de sistemas de lenguaje natural, literatura sobre sesgos algorítmicos y representación cultural, y documentos regulatorios sobre inteligencia artificial en contextos sanitarios. Se excluyeron estudios sin caracterización muestral adecuada, piezas opinativas sin fundamentación empírica y trabajos sobre tecnologías que no involucran interacción conversacional o decisiones clínicas.

La síntesis narrativa integra evidencia de distinta naturaleza mediante triangulación conceptual entre hallazgos empíricos, principios éticos y marcos jurídicos. Los casos clínicos presentados son construcciones hipotéticas basadas en patrones documentados en literatura sobre eventos adversos en salud digital y en reportes profesionales sobre uso

problemático de aplicaciones de salud mental. Su función es pedagógica e ilustrativa, no constituyen evidencia empírica de frecuencias ni pretenden demostrar causalidad.

Las limitaciones de este enfoque son explícitas. No se realizó revisión sistemática con protocolo preregistrado, no se aplicaron herramientas de evaluación de calidad metodológica estandarizadas, no se calcularon tamaños de efecto agregados ni se realizó análisis de sesgo de publicación mediante métodos cuantitativos. La validez del análisis descansa en la transparencia del proceso interpretativo, la triangulación de fuentes diversas y el contraste con marcos conceptuales establecidos en psicología clínica, bioética y derecho sanitario. La meta es ofrecer criterios operativos que permitan integrar herramientas de inteligencia artificial sin erosionar la alianza terapéutica y con salvaguardas que resguarden la seguridad del paciente.

B. LA NATURALEZA IRREDUCTIBLE DE LA RELACIÓN TERAPÉUTICA

La práctica psicológica se sostiene en un vínculo interpersonal que no se agota en la aplicación correcta de técnicas. La relación terapéutica crea un espacio de seguridad y resonancia emocional que habilita al paciente a explorar conflictos, ensayar nuevas regulaciones afectivas y reordenar significados personales. Ese espacio no es un procedimiento, sino un proceso que emerge de microajustes recíprocos a lo largo del tiempo.

Desde los primeros trabajos empíricos se observa que la calidad de la alianza terapéutica predice con consistencia los resultados clínicos. La alianza combina acuerdos sobre metas y tareas con un lazo de confianza que sostiene al paciente cuando atraviesa momentos de mayor vulnerabilidad. Esta configuración de objetivos, actividades y vínculo ofrece una base operativa para entender cómo la relación facilita el cambio psicológico y por qué su deterioro impacta de modo inmediato en la evolución del tratamiento (Bordin, 1979).

La investigación posterior confirma el papel central de este vínculo e indica que su peso explicativo suele superar el de la orientación teórica. Los estudios comparativos y las revisiones indican que la solidez de la alianza se comporta como un predictor más robusto del cambio que la técnica específica, lo que sugiere que los procesos relacionales organizan y potencian cualquier método que se utilice en sesión. La evidencia reciente subraya además que los efectos atribuidos a escuelas de pensamiento disminuyen cuando

se controla la calidad de la relación de trabajo entre terapeuta y paciente (Wampold e Imel, 2015).

También se ha mostrado que ajustar el vínculo a las necesidades de cada persona incrementa la adherencia y la probabilidad de éxito. La relación no es un molde único, sino un encuentro singular que exige sensibilidad para leer señales sutiles, flexibilidad para adaptar el encuadre y autenticidad para sostener la confianza. Por esa razón, el proceso terapéutico no puede reducirse a una secuencia estandarizada de pasos y depende de la capacidad de dos subjetividades para construir un marco compartido de sentido y de cuidado mutuo (Norcross y Lambert, 2018).

B.1. ELEMENTOS ESENCIALES DEL VÍNCULO TERAPÉUTICO

La presencia del terapeuta no se limita a estar en la sesión. Involucra atención sostenida, disponibilidad afectiva y una comprensión que incluye lo dicho y también lo que el paciente decide callar. Cuando la persona siente que su experiencia es recibida sin apresuramiento y con interés genuino, la confianza se afianza y se abre un terreno fértil para el trabajo de cambio.

La empatía en el encuadre clínico exige algo más que reconocer contenidos manifiestos. Implica acceder al mundo interno del paciente y acompañarlo desde ese marco de referencia, pero manteniendo a la vez una distancia profesional que proteja el proceso. Ese equilibrio evita la fusión y preserva la capacidad de pensar. La evidencia clásica sugiere que la combinación de aceptación incondicional, congruencia y comprensión empática aumenta la disposición al cambio y facilita la elaboración de conflictos que antes resultaban inabordables, sobre todo cuando la relación transcurre sin amenazas al sentimiento de valía personal del consultante (Rogers, 1957).

Otro componente decisivo es la contención emocional. El terapeuta recibe afectos intensos, los procesa y los devuelve en una forma que el paciente puede tolerar y pensar. De este modo, lo que inicialmente aparece como desbordante se transforma en experiencia con sentido. En contextos de trauma esta función resulta especialmente relevante porque permite que recuerdos y sensaciones intrusivas se integren de manera gradual sin reactivar circuitos de evitación o de repetición compulsiva. La contención no niega el sufrimiento. Lo sostiene el tiempo suficiente para que el paciente desarrolle recursos internos y pueda resignificar su historia con mayor estabilidad.

B.2. LA COMPLEJIDAD DEL PROCESO DIAGNÓSTICO

El diagnóstico en salud mental excede la suma de síntomas y el encaje en categorías. Supone integrar relatos de vida, patrones vinculares, hitos del desarrollo y el contexto sociocultural donde emerge el malestar. Cada manifestación adquiere sentido en una biografía concreta y cambia cuando se la ubica en la trama de experiencias del paciente. Esta lectura no solo nombra, también orienta la intervención y define prioridades clínicas.

Los sistemas artificiales trabajan con otra lógica. Detectan regularidades estadísticas en grandes conjuntos de datos y las asocian con etiquetas diagnósticas. Esa fortaleza resulta útil para la exploración inicial y para tareas de cribado, pero tiende a invisibilizar la polisemia del sufrimiento. Un mismo cuadro de signos puede responder a realidades diferentes. La ansiedad puede señalar un trastorno crónico, una fase prodrómica, una respuesta adaptativa frente a violencia o una reacción esperable ante una pérdida. Solo el juicio clínico entrenado discrimina estas posibilidades porque evalúa la función del síntoma dentro de la organización psíquica y pondera ausencias, tensiones y contradicciones del discurso.

El proceso diagnóstico es interpretativo. Requiere sensibilidad para leer matices, experiencia para sopesar hipótesis y una atención sostenida a aquello que no encaja en el relato manifiesto. Así, el diagnóstico deja de ser una etiqueta y se convierte en herramienta para decidir cómo intervenir, con qué intensidad y en qué momento. Esta perspectiva preserva la singularidad y reduce el riesgo de sobregeneralizar a partir de patrones promedio.

En ese marco, la transferencia ocupa un lugar decisivo. Describe la tendencia a reeditar en el vínculo terapéutico modelos relacionales tempranos. Lejos de ser un obstáculo, abre una vía privilegiada para comprender conflictos que se actualizan en el presente y se ponen en juego con el terapeuta. El consultante no solo cuenta lo que le ocurre. Lo pone en acto en la relación, lo que permite observar defensas, expectativas y formas de apego mientras suceden y no solo en el plano del relato. La literatura clásica ya mostraba cómo este fenómeno puede transformarse en motor de cambio cuando se reconoce y se trabaja con tacto técnico y ética del cuidado (Freud, 1912).

La transferencia es una dinámica estructural y sostenida en el tiempo. Requiere la presencia de otro real con subjetividad propia, límites, errores y reparaciones posibles.

Allí se construye un espacio donde el paciente confronta sus proyecciones, modula afectos y ensaya nuevas posiciones frente a sí y frente a los demás. Los sistemas artificiales, por más sofisticados que sean, carecen de reciprocidad y de experiencia encarnada. Pueden imitar turnos de conversación y producir respuestas coherentes, pero no participan de una relación entre sujetos. Esa ausencia limita la posibilidad de que se configure una transferencia genuina y, con ello, reduce el alcance transformador del proceso. Este límite no descalifica el uso de herramientas digitales para tareas de apoyo, pero marca con nitidez por qué no sustituyen la función clínica del terapeuta en la fase diagnóstica ni en el trabajo relacional que le sigue (Marcus, 2020).

C. EL MODELO DE NEGOCIO SUBYACENTE: LA ADICCIÓN COMO ESTRATEGIA COMERCIAL

El diseño de plataformas comerciales de inteligencia artificial para salud mental optimiza frecuencia de uso y tiempo de permanencia mediante notificaciones, recompensas intermitentes y secuencias de refuerzo. Este andamiaje replica bucles de variabilidad de recompensa estudiados en juego maquínico, donde inmediatez, imprevisibilidad y feedback parcial producen ciclos de aproximación difíciles de interrumpir (Schüll, 2012). Cuando la métrica central es tiempo en pantalla, el diseño favorece continuidad del consumo sobre consecución de objetivos clínicos.

Esta orientación contradice los fines de la psicoterapia que buscan fortalecer autonomía y promover autorregulación. Un agente digital que premia consulta constante refuerza dependencia, amplifica conductas de comprobación ansiosa y desplaza el foco desde elaboración del conflicto hacia búsqueda de alivio inmediato. En usuarios con vulnerabilidad elevada la exposición a señales de recompensa interfiere con pautas de afrontamiento que requieren pausas y reflexión.

Un uso responsable exige separar funciones clínicas de mecanismos comerciales. Las plataformas deberían informar con claridad el propósito de cada herramienta, limitar interacciones en franjas de mayor sensibilidad y ofrecer salidas hacia atención humana cuando se detectan patrones disfuncionales. La transparencia sobre incentivos y el control externo de métricas de retención son condiciones mínimas para integrar tecnología sin socavar objetivos terapéuticos.

C.1. ARQUITECTURA DE LA DEPENDENCIA DIGITAL

El diseño de estas aplicaciones prioriza la retención. Se optimizan notificaciones, rachas de uso, recompensas intermitentes y acceso inmediato para elevar la frecuencia de interacción y el tiempo de permanencia. La combinación de inmediatez y variabilidad de refuerzo activa expectativas de gratificación que se renuevan con cada consulta y transforman el malestar en un disparador de regreso al sistema. Esa dinámica entra en conflicto con el principio clínico de promover autonomía y tolerancia a la espera, ya que desalienta la pausa reflexiva y refuerza respuestas impulsivas orientadas al alivio rápido (Alter, 2017).

La lógica terapéutica propone otra cadencia. La psicoterapia trabaja con intervalos, silencios y separaciones que permiten decantar lo conversado y elaborar emociones. La meta es la independencia del paciente, no la fidelidad a un servicio. Cuando el acceso automatizado promete alivio instantáneo, el contacto continuo con la herramienta puede ocupar el lugar del trabajo de simbolización y consolidar un ciclo de consulta que reemplaza la elaboración por consumo de respuestas. Ese efecto es más visible en personas con alta sensibilidad a la gratificación inmediata o con dificultades para regular la angustia, donde la expectativa de respuesta al momento erosiona la construcción de recursos internos (Yalom, 2002).

A este cuadro se suma la potencia persuasiva de las interfaces conversacionales. Las personas tienden a responder ante computadoras como si fuesen interlocutores reales y atribuyen intención, emoción y comprensión cuando el intercambio imita rasgos humanos de forma verosímil. Un chatbot que reproduce fórmulas empáticas puede resultar convincente en situaciones de vulnerabilidad y generar la impresión de acompañamiento, aunque en realidad ofrezca una simulación consistente de diálogo sin un trasfondo relacional genuino (Reeves y Nass, 1996). La experiencia subjetiva de ser comprendido puede aparecer, pero carece de los apoyos y límites que definen un vínculo terapéutico real.

La arquitectura orientada a retención añade un sesgo adicional. Para sostener satisfacción inmediata y prolongar el uso, el sistema tiende a validar estados y creencias preexistentes del usuario en lugar de interpelarlos. Esta inclinación refuerza dinámicas de confirmación, limita la apertura a interpretaciones alternativas y reduce oportunidades de cambio. En clínica, la confrontación empática y la reencuadración de significados son

piezas decisivas para ampliar la mirada del paciente y fortalecer su autonomía. Un dispositivo que replica sin matices una lectura catastrófica del propio estado puede aliviar por un momento, pero consolida evitaciones y estancamiento. La herramienta se vuelve cómoda, no transformadora.

C.2. LA MONETIZACIÓN DEL SUFRIMIENTO PSICOLÓGICO

El modelo económico de muchas plataformas de inteligencia artificial para salud mental asigna valor a los rastros que deja cada intercambio. No se trata solo de mejorar una funcionalidad. Los registros de conversación, los vectores de emoción inferidos y las huellas de comportamiento se convierten en activos informacionales con utilidad estratégica. La captura y el procesamiento de estos datos permiten segmentar usuarios, predecir probabilidad de retorno y optimizar respuestas para aumentar la interacción. Este enfoque traslada el centro de gravedad desde la resolución del malestar hacia la producción continua de datos, una lógica estudiada en el análisis del capitalismo de la vigilancia y aplicable a servicios que dependen del seguimiento fino de la conducta digital del usuario vulnerable (Zuboff, 2019).

De esa arquitectura se desprende un incentivo estable. Las métricas que definen el éxito se concentran en tiempo de uso, frecuencia de consulta y satisfacción inmediata. El alivio rápido se vuelve un objetivo en sí mismo y desplaza el horizonte de transformación sostenida que guía la psicoterapia. Cuando la plataforma calibra respuestas para maximizar retorno, el malestar funciona como disparador de consumo. La persona vuelve a la aplicación para sentir menos angustia en el momento, pero no encuentra condiciones suficientes para procesar el conflicto de fondo. El circuito se cierra con más interacción, más datos y mayor capacidad de predicción, sin que ello garantice una mejora duradera del estado clínico.

La dimensión de privacidad agrava el cuadro. Los datos generados en contextos de salud mental describen aspectos íntimos de la vida emocional y cognitiva. Su uso con fines comerciales o su almacenamiento en entornos poco transparentes introduce riesgos de reidentificación, perfilado no deseado y circulación transfronteriza difícil de auditar. La confianza del usuario se resiente cuando no comprende quién accede a su información, con qué finalidades y durante cuánto tiempo. La literatura sobre datos médicos advierte que, aun con mecanismos de seudonimización, existen vectores de reidentificación y consecuencias sociales y económicas que exceden el ámbito individual, desde

estigmatización hasta discriminación en seguros o empleo. Por eso la gobernanza de estos datos requiere cautela, minimización y límites claros a la reutilización fuera de propósitos de cuidado y evaluación legítima de calidad (Price y Cohen, 2019).

Un uso responsable demanda transparencia sustantiva y no meramente formal. La plataforma debe informar en lenguaje claro qué datos recoge, cómo los transforma, con quién los comparte y bajo qué base jurídica se sostienen esas operaciones. También debe justificar por qué conserva información sensible durante períodos prolongados, ofrecer controles efectivos para suprimir o portar registros y someter sus prácticas a auditorías independientes. Sin estas condiciones, la promesa de ayuda se mezcla con un régimen de extracción informacional que utiliza el sufrimiento como fuente de valor y compromete la credibilidad del cuidado en salud mental.

D. LIMITACIONES FUNDAMENTALES DE LA INTELIGENCIA ARTIFICIAL EN SALUD MENTAL

Las limitaciones de la inteligencia artificial en el campo de la salud mental no se reducen a fallos de diseño o a márgenes de error estadístico. Abarcan dimensiones conceptuales que ponen en duda la posibilidad de sustituir la práctica clínica por procedimientos automatizados. En este ámbito, donde lo esencial es la comprensión del significado del sufrimiento y no solo su descripción, los sistemas artificiales enfrentan obstáculos estructurales que derivan de su propia arquitectura.

Los modelos actuales procesan correlaciones entre datos, pero carecen de intencionalidad, contexto y experiencia subjetiva. Su capacidad de detectar patrones lingüísticos o emocionales no implica comprensión del sentido que esos patrones tienen para una persona concreta. En la interacción terapéutica, una palabra, un silencio o un gesto adquieren valor según la historia y el momento vital de quien los emite. Esa interpretación requiere un marco de referencia compartido y una sensibilidad situada, elementos que los sistemas automatizados no poseen porque no participan del mundo social desde el cual se construye el significado.

El núcleo de la práctica clínica descansa en la reciprocidad: un encuentro entre dos sujetos que interpretan, se afecta y se modifican mutuamente. En ese espacio, el terapeuta se convierte en un testigo activo del proceso, capaz de reconocer ambigüedades, dudas y

contradicciones sin traducirlas de inmediato en categorías. La inteligencia artificial, por el contrario, opera desde la reducción de la ambigüedad a probabilidades. Su fortaleza es la predicción, no la comprensión, y esa diferencia marca el límite de su aplicabilidad cuando lo que está en juego no es clasificar, sino acompañar.

En consecuencia, las herramientas de inteligencia artificial pueden ofrecer apoyo instrumental. Por ejemplo, en el cribado de síntomas o en la organización de información clínica, pero no reemplazar la dimensión interpretativa, relacional y ética del encuentro terapéutico. Allí donde la intervención depende del sentido y de la presencia del otro, la sustitución tecnológica no es viable sin perder aquello que hace posible el cambio psicológico.

D.1. LA AUSENCIA DE COMPRENSIÓN SEMÁNTICA

Los modelos de lenguaje procesan patrones estadísticos en texto pero carecen de acceso al significado experiencial. Esta distinción entre manipulación de formas lingüísticas y comprensión genuina tiene consecuencias clínicas directas que no pueden ignorarse. La investigación reciente demuestra que estos sistemas captan correlaciones superficiales sin construir representaciones semánticas estables, un hallazgo que se confirma en evaluaciones de robustez donde aparecen fallos sistemáticos ante variaciones contextuales mínimas (Bender y Koller, 2020; Ribeiro et al., 2020).

En psicoterapia esta limitación se vuelve crítica. Un paciente latinoamericano que dice tener el corazón apretado expresa angustia emocional mediante una metáfora cultural, pero un sistema detecta las palabras corazón y apretado y puede sugerir evaluación cardíaca. Una adolescente que responde estoy perfectamente bien con tono sarcástico tras pelear con sus padres comunica malestar intenso, aunque el chatbot registre un estado positivo. Un paciente traumatizado evita mencionar el abuso directamente, usando eufemismos y pausas prolongadas que no indican riesgo en el nivel literal pero sí en el patrón comunicativo total. Estas situaciones no son excepcionales sino constitutivas del trabajo clínico cotidiano.

La práctica terapéutica exige integrar contenido literal, prosodia, contexto biográfico e historia relacional para construir significado. Los sistemas automatizados operan sin intencionalidad ni experiencia encarnada, lo que genera el problema de fundamentación simbólica donde los símbolos no están anclados en experiencia del mundo (Harnad,

1990). Esta brecha no se cierra con más datos ni con arquitecturas más complejas porque responde a una limitación ontológica de base.

La implicancia práctica es clara. Las herramientas de inteligencia artificial pueden apoyar documentación, detectar palabras clave de riesgo y organizar información longitudinal. Sin embargo, la interpretación semántica contextualizada que resulta esencial para diagnóstico diferencial, evaluación de riesgo y construcción de alianza permanece bajo responsabilidad clínica exclusiva. La automatización sin supervisión profesional en estas funciones constituye un riesgo de seguridad documentado y previsible.

D.2. INCAPACIDAD PARA LA DETECCIÓN DE COMUNICACIONES ENGAÑOSAS

Un límite persistente de los sistemas automatizados es su dificultad para reconocer distorsiones, omisiones o engaños en el relato del paciente. Las personas tampoco son infalibles, pero combinan canales diversos que incrementan la probabilidad de acierto. Integran contenido, prosodia, pausas, postura, miradas y la evolución del vínculo en el tiempo. Esa lectura multimodal permite ajustar hipótesis, sostener la duda cuando corresponde y revisar interpretaciones a la luz de nuevas señales contextuales (Bond y DePaulo, 2006).

Los sistemas basados en lenguaje operan casi exclusivamente sobre el texto o la transcripción. Carecen de acceso estable y confiable a pistas finas como microvariaciones faciales, cambios súbitos de tono, latencias inusuales de respuesta o desajustes entre gesto y enunciado. En la clínica, estos indicios son decisivos para detectar incongruencias entre lo que se dice y lo que se siente. La literatura describe cómo microexpresiones, modulaciones y quiebres del discurso pueden revelar tensiones ocultas que no aparecen en el contenido literal, especialmente cuando la persona intenta minimizar, disimular o probar límites del encuadre terapéutico (Ekman, 2001).

La limitación se vuelve crítica en escenarios de riesgo. En la ideación suicida, muchos pacientes comunican malestar de forma indirecta, con alusiones, eufemismos o giros que evitan la mención explícita del plan. La valoración clínica no se apoya solo en palabras clave, sino en patrones de ambivalencia, vacíos en el relato, oscilaciones afectivas y señales de desesperanza que emergen en la interacción. Un sistema orientado a correspondencias semánticas puede pasar por alto estas configuraciones y ofrecer

respuestas tranquilizadoras cuando se necesita escalado y contención inmediata. La consecuencia es un aumento del riesgo por falsa seguridad y por demora en la derivación (Joiner, 2005).

Reconocer este límite no implica renunciar a las herramientas digitales. Señala el campo donde su aporte debe ser secundario y supervisado. La detección de comunicaciones engañosas requiere sensibilidad situacional, contraste con el historial del vínculo y una integración de señales que hoy excede el alcance de los modelos centrados en texto. La función responsable de la tecnología en estos casos es apoyar el registro y la organización de información, no sustituir el juicio clínico que decide cuándo preguntar más, cuándo dudar y cuándo intervenir.

D.3. SESGOS ALGORÍTMICOS Y REPRESENTACIÓN CULTURAL

Los sistemas de inteligencia artificial reproducen los sesgos presentes en los datos de entrenamiento. El estudio que documentó disparidades en clasificadores comerciales de rostro mostró tasas de error muy superiores en mujeres de piel oscura respecto de varones blancos y desarmó la idea de neutralidad automática de las máquinas. La lección es directa. Cuando un conjunto de entrenamiento no representa con equilibrio a los grupos poblacionales, el modelo aprende reglas que funcionan mejor para quienes están sobrerrepresentados y falla con mayor frecuencia en quienes apenas figuran en los datos de origen (Buolamwini y Gebru, 2018).

El problema excede el reconocimiento facial. La forma en que los sistemas indexan y ordenan información puede reforzar estereotipos y empujar a los usuarios hacia asociaciones sesgadas que circulan en el espacio público. Un enfoque basado en correlaciones no distingue entre patrones útiles y patrones injustos si ambos optimizan una métrica de rendimiento. El resultado puede ser la amplificación de prejuicios latentes en los corpus utilizados para entrenar y evaluar modelos, con efectos concretos sobre cómo se perciben y tratan colectivos completos de personas en ámbitos sensibles como la salud (Noble, 2018).

En salud mental, estos desajustes tienen consecuencias clínicas. La investigación transcultural muestra que síntomas, explicaciones del malestar y estrategias de afrontamiento dependen del contexto. Lo que en una cultura aparece como un signo esperado de duelo, en otra puede leerse como depresión. Lo que en un marco comunitario

opera como práctica espiritual con función reguladora, en un modelo universalizado puede figurar como pensamiento disfuncional. Comprender el sufrimiento requiere categorías locales y atención a los significados que las personas asignan a su experiencia, no solo coincidencias con un catálogo de etiquetas (Kleinman, 1988). La práctica culturalmente sensible ofrece un criterio de corrección. Examina supuestos, reconoce el peso de las categorías del clínico y adapta evaluación e intervención a biografías y entornos específicos. Esta orientación reduce diagnósticos erróneos y evita patologizar conductas que cumplen funciones adaptativas. Cuando los datos de entrenamiento provienen de contextos mayoritariamente occidentales, urbanos y con niveles educativos particulares, crece la probabilidad de malinterpretar expresiones culturales de malestar. Sin estrategias de inclusión de datos y validaciones en poblaciones diversas, la inteligencia artificial tiende a universalizar lo particular y a reproducir inequidades en acceso y calidad de la atención psicológica (Sue y Sue, 2015).

Este riesgo se intensifica en Latinoamérica por un triple desafío lingüístico, cultural y contextual. En primer lugar, evaluaciones comparativas han señalado que los grandes modelos de lenguaje muestran menor desempeño cuando operan en español en dominios clínicos, especialmente al manejar terminología técnica o matices pragmáticos propios del registro sanitario. En segundo lugar, la expresión del malestar está atravesada por categorías culturales como nervios, susto o ataque de nervios, que condensan experiencias de sufrimiento difíciles de mapear de manera directa a diagnósticos estandarizados. Un sistema carente de sensibilidad cultural puede patologizar respuestas socialmente aceptadas al estrés o, a la inversa, minimizar síntomas que exigen intervención. Además, rasgos relacionales como el familismo o el personalismo organizan expectativas de cuidado y la alianza terapéutica, dimensiones que un intercambio automatizado no reproduce con fidelidad. En tercer lugar, el contexto sociopolítico regional, marcado por desigualdad, violencia y desplazamientos, introduce determinantes sociales del malestar que exigen lectura situada. Un modelo no entrenado para reconocer estos factores diluye su análisis en explicaciones individualizantes y pierde de vista fuentes estructurales de sufrimiento. Integrar datos locales, evaluar desempeño por subpoblaciones y co-diseñar herramientas con profesionales y comunidades de la región no es un complemento opcional, es una condición para que la tecnología no agrave brechas ya existentes.

E. RIESGOS ESPECÍFICOS PARA POBLACIONES VULNERABLES

El impacto de sistemas de inteligencia artificial en salud mental no es uniforme. Algunas poblaciones presentan vulnerabilidad aumentada por etapa de desarrollo, limitaciones en capacidad de juicio o naturaleza de sus padecimientos. Sustituir intervención humana por automatización puede agravar el malestar y generar daños difíciles de revertir.

En infancia y adolescencia el proceso terapéutico estructura identidad, consolida regulación emocional y ensaya modos seguros de vinculación. La interacción con agentes artificiales carece de reciprocidad auténtica y no provee modelado emocional que permite internalizar límites. Donde la simbolización está en desarrollo, las respuestas estandarizadas generan confusión sobre naturaleza de las relaciones.

En personas mayores la sustitución tecnológica refuerza aislamiento y erosiona redes de apoyo esenciales. La dependencia de asistentes conversacionales oculta signos de deterioro cognitivo o depresión, retrasando detección clínica y acceso a tratamientos adecuados.

Las personas con trastornos graves de personalidad o trauma complejo necesitan relación terapéutica estable donde coherencia, presencia y reparación de rupturas sean componentes centrales. Los sistemas que no perciben ambivalencia emocional intensifican sentimientos de vacío y privan al paciente de oportunidades para reelaborar vínculos.

Las poblaciones con barreras lingüísticas, discapacidad o acceso limitado a recursos digitales enfrentan doble afectación. La estandarización cultural incrementa errores de interpretación mientras la dependencia de sistemas automatizados reproduce desigualdades previas e introduce nuevas formas de exclusión bajo apariencia de accesibilidad tecnológica.

La incorporación de inteligencia artificial exige enfoque diferenciado que analice impacto sobre grupos cuya vulnerabilidad hace imprescindible supervisión humana constante.

E.1. IMPACTO EN ADOLESCENTES Y ADULTOS JÓVENES

La adolescencia es una etapa decisiva para consolidar identidad y regular emociones. El desarrollo neurológico y social necesita vínculos humanos confiables que ofrezcan modelos de relación, sostén y validación. La seguridad afectiva que brindan esos lazos

permite experimentar, equivocarse y reparar sin pérdida de estima, condiciones que vuelven posible integrar una imagen coherente de sí mismo y del propio mundo relacional (Steinberg, 2013; Erikson, 1968).

El uso permanente de sistemas automatizados puede interferir en esos aprendizajes. La evidencia sobre generaciones altamente expuestas a tecnologías digitales describe mayores dificultades para tolerar la frustración y para construir estrategias internas de afrontamiento, con señales de aumento de malestar y de dependencia de alivios rápidos. Estos hallazgos no son uniformes en todos los contextos, pero invitan a cautela cuando la herramienta digital ocupa el lugar del acompañamiento humano y del trabajo reflexivo que requieren los conflictos propios de la edad (Twenge, 2017). Si un adolescente recurre de modo sistemático a un chatbot para modular emociones intensas, corre el riesgo de consolidar un circuito de consulta inmediata que posterga la elaboración y empobrece los recursos para la vida adulta.

La preocupación aumenta ante escenarios de riesgo clínico. La detección precoz de ideación suicida depende a menudo de señales indirectas, cambios sutiles de tono y variaciones en la conducta que emergen en el contacto humano sostenido. La valoración clínica integra ambivalencias, silencios y contradicciones que no siempre se expresan en palabras literales. Un sistema centrado en patrones lingüísticos explícitos puede normalizar giros peligrosos o pasar por alto indicios encubiertos, con demoras en la derivación y en el acceso a ayuda especializada. En población juvenil, ese retraso puede empeorar el pronóstico y reducir oportunidades de intervención preventiva eficaz (Gould, Greenberg, Velting y Shaffer, 2003).

E.2. PERSONAS CON TRASTORNOS DE PERSONALIDAD

Los trastornos de personalidad se reconocen por patrones persistentes de experiencia interna y conducta que se apartan de las expectativas culturales y comprometen la cognición, la afectividad, el control de impulsos y el funcionamiento interpersonal. No se trata solo de rasgos intensos, sino de modos estables de percibir y vincularse que generan sufrimiento y deterioro en áreas significativas de la vida (American Psychiatric Association, 2013).

En el trastorno límite de la personalidad la oscilación entre idealización y devaluación organiza gran parte de la vida relacional. La disponibilidad continua de un chatbot puede

amplificar ese vaivén. En un primer momento el sistema puede ser investido como una presencia perfecta que responde siempre y al instante. Ante la primera falla, demora o respuesta percibida como poco empática, sobreviene la devaluación tajante. La lógica de acceso inmediato alimenta fantasías de control y dificulta la construcción de tolerancia a la frustración y de permanencia de objeto, aprendizajes que son centrales para estabilizar los vínculos. El trabajo clínico efectivo con este cuadro requiere límites claros, validación cuidadosa y entrenamiento en regulación emocional dentro de una relación consistente y reparadora, condiciones que exceden lo que puede ofrecer una interacción automatizada sin subjetividad ni memoria relacional genuina (Linehan, 1993).

Los cuadros narcisistas presentan otro desafío. Las defensas grandiosas funcionan como sostén de un sentido de sí mismo frágil y expuesto. El proceso terapéutico demanda una confrontación empática que señale distorsiones sin humillar, acompañe la vergüenza y promueva formas más realistas de autoestima. Un sistema artificial que responde de modo incondicional y provee validaciones constantes puede transformarse en una fuente de gratificación narcisista que mantiene la escisión entre grandiosidad y vacío. Falta la presencia de un otro humano capaz de modular la intensidad, reconocer microfracturas del vínculo y trabajar las rupturas y reparaciones que organizan el cambio en este terreno clínico (Kernberg, 1975; Kohut, 1971).

En ambos grupos el riesgo de sustitución tecnológica es alto. La intervención requiere una relación viva donde la ambivalencia pueda ser nombrada y procesada, donde se negocien límites y se trabaje la continuidad del lazo a lo largo del tiempo. La arquitectura de los agentes conversacionales favorece el alivio inmediato y la repetición de patrones disfuncionales de búsqueda de confirmación. Sin encuadre humano, la herramienta tiende a reforzar aquello que se intenta transformar. Por esa razón su uso solo resulta prudente como apoyo complementario y bajo supervisión clínica estrecha, con criterios de derivación claros y con objetivos acotados que no invadan el núcleo del trabajo relacional.

E.3. VÍCTIMAS DE TRAUMA COMPLEJO

El trauma complejo, producto de experiencias prolongadas de abuso, negligencia o violencia, altera la organización psíquica y dificulta el establecimiento de confianza con los otros. No solo deja síntomas, también modifica la forma en que la persona evalúa el peligro, procesa la emoción y se vincula. La literatura clínica describe patrones de hipervigilancia, desregulación afectiva, disociación y dificultades para sostener la

continuidad del yo, todo ello atravesado por expectativas relacionales marcadas por el daño y la traición de la confianza básica (Herman, 1992).

Desde una perspectiva neurofisiológica, la teoría polivagal ayuda a entender por qué estos estados se perpetúan. La evaluación automática de seguridad o amenaza organiza el acceso a circuitos de calma, lucha o desconexión. En muchos sobrevivientes la señal de peligro se activa con facilidad y de forma sostenida, lo que limita la capacidad de mentalizar y reduce la ventana de tolerancia para el trabajo emocional. La alternancia entre hiperactivación y colapso no es simplemente un estado de ánimo, sino una respuesta corporal aprendida que condiciona la posibilidad de participar en una conversación terapéutica sin desbordarse ni desconectarse (Porges, 2011).

Por estas razones el tratamiento requiere la presencia de un terapeuta que pueda co-regular emociones, graduar la intensidad de las intervenciones y ofrecer un entorno predecible. La terapia efectiva se planifica por fases. Primero se prioriza estabilización y seguridad, luego se trabaja integración de memorias y finalmente se consolidan vínculos y proyectos vitales. Si la exposición a contenidos traumáticos se realiza en un momento inadecuado o con una intensidad excesiva, se reactivan recuerdos fragmentados, aumenta la disociación y se refuerzan circuitos de defensa que ya eran desadaptativos. La clínica del trauma complejo muestra que el tempo de la intervención y la capacidad de sostener rupturas y reparaciones en la relación son componentes técnicos ineludibles, imposibles de delegar en un sistema que responde por patrones lingüísticos sin sensibilidad situacional encarnada (van der Hart, Nijenhuis y Steele, 2006).

La reconstrucción de la confianza depende de experiencias relacionales reparadoras. Se avanza cuando la persona comprueba en el vínculo presente que su palabra es creída, que los límites se respetan y que los momentos de malentendido pueden repararse sin castigo ni abandono. Esa vivencia configura una base segura desde la cual explorar emociones y ensayar nuevas formas de apego. Un sistema artificial puede ofrecer disponibilidad constante y lenguaje de apoyo, pero carece de la autenticidad, la imprevisibilidad contenida y la responsabilidad mutua que definen un vínculo real. Sin esa base, el aprendizaje relacional que posibilita cambios duraderos queda truncado y el riesgo de retraumatización aumenta por ausencia de co-regulación humana y de un encuadre capaz de sostener la complejidad del trauma (Bowlby, 1988).

F. IMPLICANCIAS ÉTICAS Y DEONTOLÓGICAS

La incorporación de sistemas de inteligencia artificial en salud mental exige releer principios clásicos de bioética y deontología profesional a la luz de interacción mediada por tecnología. Autonomía, beneficencia, no maleficencia y justicia siguen siendo el eje, pero su aplicación demanda criterios operativos que contemplen sesgos algorítmicos, opacidades técnicas y asimetrías informativas.

Respetar la autonomía requiere consentimiento informado específico que detalle funciones, límites, riesgos previsibles, tratamiento de datos y vías de derivación clínica. Este consentimiento debe ser comprensible, verificable y revocable sin penalizaciones, ya que la decisión de usar o no un agente automatizado forma parte de la autodeterminación del paciente en contextos de vulnerabilidad (Beauchamp y Childress, 2019; APA, 2017).

Beneficencia y no maleficencia exigen evidencia de eficacia y seguridad. La prudencia ética implica validar usos concretos, fijar umbrales de escalado a atención humana y monitorizar efectos adversos como incremento de ansiedad o retraso en búsqueda de ayuda profesional. Cuando un sistema no explica suficientemente por qué sugiere una respuesta en casos sensibles, la obligación de cuidado inclina la balanza hacia intervención humana y protocolos de auditoría que documenten trayectorias de decisión (Beauchamp y Childress, 2019; UNESCO, 2021).

El principio de justicia obliga a evitar reproducción de desigualdades. La ética profesional demanda evaluar sesgos de datos por subpoblaciones, ajustar el diseño para no excluir usuarios con barreras lingüísticas o tecnológicas y garantizar accesibilidad sin degradar calidad. La justicia también se expresa en gobernanza de datos mediante minimización, propósito limitado y controles de acceso estrictos. La reutilización con fines comerciales o transferencia opaca entre jurisdicciones vulnera expectativas legítimas de confidencialidad (APA, 2017; UNESCO, 2021).

La deontología añade obligaciones específicas. El principio de competencia impone formarse para comprender alcances y límites antes de integrar herramientas en la práctica. La honestidad profesional prohíbe antropomorfizar sistemas o prometer prestaciones que no pueden ofrecer. El deber de advertencia exige diseñar rutas claras para intervención en crisis cuando el contenido sugiere riesgo inminente. La responsabilidad es indelegable

y requiere acuerdos explícitos de supervisión, registro y auditoría que aseguren trazabilidad (APA, 2017; Beauchamp y Childress, 2019; UNESCO, 2021).

F.1. AUTONOMÍA E INFORMACIÓN VERAZ

El respeto a la autonomía exige que las personas comprendan con claridad qué servicio reciben y que cuenten con información suficiente para decidir libremente. El consentimiento informado no es un formulario, es un proceso continuo que acompaña la toma de decisiones y se renueva cuando cambian las funciones, los riesgos o los usos de los datos. Esa es la concepción extendida en la teoría y en la práctica biomédica moderna, que entiende el consentimiento como un intercambio comprensible, voluntario y revocable sin penalizaciones (Faden y Beauchamp, 1986).

En aplicaciones de salud mental esta condición se incumple con frecuencia. Muchos usuarios no saben que interactúan con sistemas automatizados que simulan empatía sin intervención humana directa. Esa omisión vulnera la transparencia clínica y debilita la autodeterminación, ya que impide valorar con realismo capacidades, límites y riesgos del agente. El problema se agrava cuando la plataforma se presenta como alternativa equivalente a la psicoterapia profesional y refuerza expectativas que no puede sostener. La literatura sobre ética de la inteligencia artificial advierte que los principios generales no bastan si no se acompañan de mecanismos verificables de implementación, auditoría y comunicación clara orientada al usuario no experto (Mittelstadt, 2019).

Un enfoque responsable requiere requisitos mínimos de información antes del primer uso y durante todo el ciclo de interacción. Se deben explicitar funciones reales del sistema, alcances y límites de uso, criterios de derivación a atención humana, tratamiento y conservación de datos, actores que acceden a la información y bases jurídicas que habilitan esas operaciones. También se deben ofrecer controles efectivos para retirar el consentimiento, suprimir registros y obtener copias legibles de las interacciones. Las revisiones comparativas de guías globales señalan que la falta de transparencia en estos puntos es una de las carencias más persistentes, por lo que conviene estandarizar formatos de lectura fácil y evidencias de comprensión, además de habilitar auditorías independientes de la información brindada al público (Jobin, Ienca y Vayena, 2019).

F.2. BENEFICENCIA Y NO MALEFICENCIA

Los principios de beneficencia y no maleficencia obligan a demostrar que una intervención maximiza beneficios y reduce riesgos. No basta con la promesa de accesibilidad o de alivio inmediato. En salud mental se exige evidencia suficiente de eficacia y, sobre todo, de seguridad, porque intervenciones aparentemente inocuas pueden producir efectos adversos difíciles de detectar si no se los monitoriza con método. Esta es la interpretación dominante en la ética biomédica contemporánea, que sitúa el deber de cuidado por encima de ventajas comerciales o de conveniencia tecnológica y reclama justificar cada uso con datos y con procedimientos de control proporcionales al riesgo (Beauchamp y Childress, 2019).

La cuestión central es si los beneficios superan los riesgos en condiciones reales. La evidencia disponible sobre chatbots y agentes conversacionales es todavía limitada y heterogénea. Predominan medidas de satisfacción inmediata y autorreportes de corto plazo, con tamaños muestrales pequeños y seguimientos breves. Falta conocer con rigor si los efectos se sostienen en el tiempo y si la herramienta reduce síntomas sin generar dependencia, postergación de consulta presencial o incremento de conductas de comprobación ansiosa. La literatura previa sobre intervenciones digitales ya advertía que la calidad metodológica y la claridad en la definición de resultados constituyen cuellos de botella para traducir hallazgos en recomendaciones clínicas robustas, sobre todo cuando la intervención se apoya en interacción conversacional y no en módulos psicoeducativos estructurados (Mohr, Burns, Schueller, Clarke y Klinkman, 2013).

En contextos de incertidumbre la orientación prudencial es clara. El principio de precaución, recogido en marcos de gobernanza internacional de la inteligencia artificial, sugiere limitar despliegues amplios cuando faltan evaluaciones de seguridad y mecanismos independientes de auditoría. Aplicado a salud mental, esto implica validar usos concretos, fijar umbrales de escalado a atención humana, registrar eventos adversos y publicar resultados, incluidos los negativos. Sin estudios longitudinales que describan efectos sostenidos, cualquier expansión masiva corre el riesgo de consolidar prácticas que alivian a corto plazo, pero erosionan autonomía y retrasan la búsqueda de ayuda profesional. La ética de la implementación demanda, por tanto, pilotos acotados, métricas clínicas definidas a priori y transparencia sobre desempeño y fallos, en lugar de

suposiciones optimistas basadas en aceptación inicial o en métricas de interacción (UNESCO, 2021; Beauchamp y Childress, 2019).

F.3. JUSTICIA Y EQUIDAD EN EL ACCESO

El argumento más recurrente a favor de la inteligencia artificial en salud mental es su capacidad para ampliar cobertura y reducir barreras. En teoría, las plataformas digitales acortan distancias, alivian listas de espera y ofrecen apoyo básico donde hoy no hay alternativas. Sin embargo, la experiencia acumulada indica que la incorporación de tecnología no garantiza por sí sola una disminución de las desigualdades; cuando no se acompaña de políticas activas de equidad puede reproducirlas o incluso profundizarlas. El acceso desigual a dispositivos adecuados, conectividad estable y alfabetización digital excluye a amplios sectores bajo la apariencia de modernización (Vayena, Dzenowagis, Brownstein y Sheikh, 2018).

El riesgo ético aumenta cuando estas herramientas se presentan como sustitutos de la atención profesional. En contextos con recursos limitados, la tentación de ofrecer sistemas automatizados como única opción para poblaciones vulnerables consolida una atención de segunda categoría. Las personas con menos recursos terminan recibiendo asistencia algorítmica mientras la atención presencial y especializada queda reservada para quienes pueden costearla. Así se naturaliza un sistema dual que erosiona el principio de justicia distributiva que orienta la política sanitaria.

La literatura en salud pública muestra que los determinantes sociales, como ingreso, educación, entorno y redes de apoyo, condicionan de manera decisiva las oportunidades de cuidado (Marmot, 2005). Desde esta perspectiva, la justicia sanitaria no se agota en la disponibilidad de servicios. Exige equidad en la calidad de las intervenciones y en la capacidad efectiva de las personas para beneficiarse de ellas. Aplicado a la inteligencia artificial, esto supone evaluar el impacto diferencial por nivel socioeconómico, idioma, edad y región, e introducir medidas correctivas cuando se detecten brechas. De otro modo, el despliegue tecnológico sin compensaciones podría legitimar soluciones de menor valor para quienes más protección necesitan (Daniels, 2007).

En América Latina este debate requiere un escrutinio aún más cuidadoso. La combinación de financiamiento público limitado en salud mental, desigualdades territoriales persistentes, conectividad irregular y diversidad lingüístico-cultural crea condiciones en

las que la promesa de “democratizar el acceso” puede derivar en sustitución de bajo costo. Zonas rurales, comunidades indígenas y barrios con infraestructura digital precaria corren el riesgo de quedar relegados a chatbots y aplicaciones autónomas, mientras los servicios presenciales se concentran en centros urbanos y en usuarios con mayor capacidad de pago. Lejos de cerrar brechas, un despliegue acrítico podría consagrar un circuito de atención de inferior calidad para quienes ya enfrentan mayores barreras. La integración responsable exige salvaguardas explícitas: acompañamiento profesional, evaluación de impacto por subpoblaciones, accesibilidad lingüística y cultural, y mecanismos de derivación oportuna a atención humana cuando la complejidad clínica lo demande.

G. CONSIDERACIONES JURÍDICAS Y REGULATORIAS

El despliegue de sistemas de inteligencia artificial en salud mental exige vincular obligaciones al uso real de cada herramienta. Cuando el sistema participa en decisiones clínicas o realiza triaje que condiciona el acceso a tratamiento, las exigencias se elevan y requieren validación previa en población pertinente, supervisión humana efectiva y vigilancia posterior a la comercialización. Esta lógica se alinea con los marcos que ordenan requisitos de manera proporcional al riesgo y que exigen trazabilidad técnica, documentación suficiente para reconstruir decisiones y planes explícitos de supervisión durante todo el ciclo de vida del sistema.

El enfoque graduado por riesgo que avanza en Europa ofrece un punto de referencia útil. Establece gobernanza de datos para mitigar sesgos, registros de eventos para auditorías independientes y evaluación antes de la puesta en el mercado, junto con mecanismos de vigilancia continuada para detectar efectos adversos y activar correcciones. En ese esquema, los sistemas de información de bajo impacto operan con transparencia reforzada, mientras que las aplicaciones que inciden en diagnóstico o tratamiento se someten a evaluaciones clínicas, gestión de riesgos centrada en seguridad del paciente y control de rendimiento en condiciones reales. Las afirmaciones comerciales sobre eficacia deben contar con respaldo empírico y pierden legitimidad cuando sugieren sustitución profesional sin salvaguardas o cuando exageran resultados sin pruebas suficientes.

El tratamiento de datos personales en salud demanda un estándar elevado. La información producida por estas aplicaciones incluye datos sensibles y también inferencias sobre estados mentales, lo que obliga a licitud clara, minimización, limitación de finalidad y derechos efectivos de acceso, rectificación y supresión. El consentimiento ha de ser informado, específico y revocable sin penalización. La reutilización con fines comerciales ajenos al cuidado contradice expectativas razonables y erosiona la confianza. Además, la transferencia transfronteriza de datos debe evaluarse por su impacto en la protección efectiva de las personas usuarias y por la capacidad real de hacer valer derechos en jurisdicciones de destino.

La responsabilidad por daños no se agota en el contrato. Los regímenes de responsabilidad por productos y por servicios pueden activarse cuando un defecto de diseño, entrenamiento o despliegue produce perjuicio. La opacidad técnica no exime del deber de diligencia. Los proveedores deben prever rutas de auditoría, acceso a registros, explicación comprensible de resultados clínicamente relevantes y mecanismos de corrección verificables. Las instituciones sanitarias que integran estas herramientas mantienen deberes de selección, supervisión y monitoreo y no pueden delegar la responsabilidad clínica en un algoritmo.

El panorama latinoamericano añade desafíos de implementación. La región cuenta con marcos generales de protección de datos y con normas sanitarias que resguardan el secreto profesional, pero rara vez existen disposiciones específicas para sistemas de inteligencia artificial que inciden en decisiones clínicas. Esta brecha genera zonas grises en clasificación de riesgos, validación local, reporte de eventos adversos y control de publicidad. Para evitar un trasplante normativo acrítico conviene adaptar los principios de proporcionalidad y transparencia a las realidades de infraestructura, conectividad y diversidad cultural de la región, con participación efectiva de autoridades sanitarias, colegios profesionales y organizaciones de pacientes.

En Argentina el andamiaje jurídico ofrece puntos de apoyo que permiten avanzar hacia una integración responsable. La normativa de protección de datos reconoce la categoría de datos sensibles y la necesidad de consentimiento válido, a la vez que el secreto profesional en salud obliga a una custodia reforzada de historias clínicas y comunicaciones terapéuticas. La autoridad sanitaria nacional regula los productos médicos y puede abarcar software con finalidad de diagnóstico o apoyo a decisiones clínicas cuando su uso impacta la seguridad del paciente. En ese supuesto se justifican

requisitos de registro, evidencia clínica acorde al uso previsto, gestión de riesgos y vigilancia luego de su comercialización. Las afirmaciones de marketing sobre beneficios terapéuticos deben ser veraces y verificables y no pueden inducir a confusión respecto de la supervisión profesional. La adopción institucional de estas herramientas requiere además políticas claras de consentimiento informado, evaluación de impacto en protección de datos, registros de auditoría y protocolos de derivación inmediata a atención humana ante señales de riesgo.

Una regulación responsable asigna obligaciones según función, demuestra seguridad con evidencia suficiente, protege datos con estándares elevados y mantiene la primacía de la supervisión humana cuando la intervención afecta la salud. La coordinación entre autoridades de protección de datos, reguladores sanitarios y colegios profesionales permitirá traducir estos principios en procedimientos auditables y en incentivos correctos para que la innovación digital complemente la práctica clínica sin desvirtuar su núcleo humano.

G.1. RESPONSABILIDAD CIVIL Y PENAL

Atribuir responsabilidad cuando un sistema de inteligencia artificial produce un daño en el contexto terapéutico es una de las cuestiones más complejas del derecho contemporáneo. Los modelos jurídicos tradicionales de responsabilidad objetiva por producto defectuoso y negligencia profesional por prestación inadecuada de servicios fueron diseñados para entornos en los que podía identificarse un agente humano o una cadena de decisiones rastreable. En cambio, los sistemas autónomos operan con cierto grado de imprevisibilidad, generan resultados que no siempre son reproducibles y se apoyan en modelos opacos, lo que dificulta establecer culpa, causalidad y previsibilidad en sentido clásico (Abbott, 2020).

El primer obstáculo es de naturaleza conceptual. Todavía no está claro si quien utiliza un chatbot terapéutico debe considerarse paciente, consumidor o usuario de software. Cada categoría activa regímenes de protección diferentes. El tratamiento como paciente hace regir deberes de diligencia profesional y de protección sanitaria. La consideración como consumidor aplica normas de protección contractual y de información. La asimilación a usuario de software reduce el vínculo a términos de licencia privados con menor control público (Levi, 2018). Esta ambigüedad genera un vacío jurídico que deja a los afectados en una zona gris donde los mecanismos de reparación son inciertos o inaccesibles.

El problema se amplía por la práctica extendida de incluir cláusulas de exención de responsabilidad en los términos de uso. Estas disposiciones habituales en plataformas tecnológicas intentan limitar la exposición de los desarrolladores ante eventuales reclamos, pero resultan incompatibles con el principio de protección del usuario en contextos de salud. En la práctica dichas cláusulas trasladan todo el riesgo al individuo y reducen su capacidad de reclamar por daños derivados de fallos de diseño, entrenamiento deficiente o falta de supervisión humana. Además, la opacidad de los algoritmos agravada por el secreto comercial impide determinar si el error proviene de sesgos en los datos, defectos en la arquitectura o negligencia en la actualización de modelos (Pasquale, 2015).

La consecuencia es un escenario donde las víctimas enfrentan perjuicios sin saber contra quién accionar ni bajo qué régimen. La indefinición no solo debilita la confianza en las tecnologías de salud mental sino que erosiona la eficacia preventiva del derecho. Frente a ello resulta necesario establecer marcos regulatorios específicos que delimiten responsabilidades según el rol del desarrollador, proveedor, integrador o profesional que supervisa la herramienta, y que impongan obligaciones de transparencia, trazabilidad y aseguramiento. Tales medidas permitirían reconstruir la cadena causal y ofrecer mecanismos de compensación efectivos, preservando al mismo tiempo la innovación tecnológica dentro de límites jurídicos claros y previsibles.

G.2. PROTECCIÓN DE DATOS Y CONFIDENCIALIDAD

El tratamiento de información psicológica sensible abre un frente de dilemas que no se resuelve con cláusulas genéricas de privacidad. La ética de los datos busca proteger dignidad y autonomía, no solo custodiar archivos. Esto exige preguntarse qué se recoge, para qué se usa, quién accede y con qué límites, así como reconocer que ciertos usos pueden ser indebidos aun cuando resulten técnicamente posibles o legalmente ambiguos (Floridi y Taddeo, 2016).

En la relación clínica tradicional el secreto profesional establece una barrera nítida. En la interacción con sistemas automatizados esa barrera se diluye y queda sujeta a políticas de plataforma y a marcos generales de protección de datos que rara vez contemplan la especificidad de la salud mental. Incluso normativas consolidadas en contextos de referencia han mostrado dificultades para garantizar confidencialidad efectiva cuando el cuidado se mediatiza por tecnología y los flujos informacionales se vuelven más complejos que los supuestos por la regulación original (Gostin y Nass, 2009).

La recopilación masiva y el análisis secundario añaden riesgos. No se almacenan solo las interacciones inmediatas. Se derivan perfiles psicológicos, se infieren rasgos y patrones conductuales y se construyen historiales que pueden reutilizarse con finalidades ajenas al cuidado, desde segmentación comercial hasta experimentación de producto. Estas prácticas expanden la superficie de exposición, aumentan la posibilidad de reidentificación y desplazan el control desde el paciente hacia el proveedor tecnológico, con consecuencias que pueden afectar acceso a servicios, asegurabilidad o empleo (Vayena, Salathé, Madoff y Brownstein, 2015).

El consentimiento es un punto crítico. La aceptación de políticas extensas, técnicas y poco claras conduce a decisiones sin comprensión real del alcance. La ilusión de control individual fracasa cuando los usuarios no pueden valorar qué se comparte, durante cuánto tiempo y con quién. El dilema se agrava ante transferencias internacionales a jurisdicciones con salvaguardas débiles, donde el ejercicio de derechos de acceso, rectificación o supresión se vuelve incierto o impracticable. En este escenario, apelar solo a la autogestión de la privacidad resulta insuficiente para proteger intereses legítimos y expectativas razonables de confidencialidad en salud mental (Solove, 2013).

Un uso responsable requiere minimización estricta de datos, finalidades delimitadas, ciclos de conservación acotados y auditorías independientes sobre acceso y reutilización. También demanda formatos de información comprensibles, evidencias de entendimiento por parte del usuario y mecanismos efectivos para revocar consentimiento sin penalizaciones. Sin estas garantías, la promesa de ayuda se entrecruza con un régimen de extracción informacional que erosiona la confianza y compromete el núcleo ético del cuidado psicológico.

G.3. EJERCICIO ILEGAL DE LA PSICOLOGÍA

Un tercer problema emerge al considerar la legalidad del ejercicio profesional. En numerosos países la práctica de la psicología es una actividad regulada y reservada a quienes acreditan formación, habilitación y sujeción a códigos de conducta. Este cerco protege a la población frente a intervenciones no cualificadas y fija obligaciones de competencia, confidencialidad, evaluación continua del riesgo y supervisión clínica. Cuando un sistema de inteligencia artificial emite diagnósticos, formula hipótesis psicopatológicas o sugiere planes de tratamiento, replica actos propios de la profesión sin

satisfacer los estándares de formación, responsabilidad y control externo que la regulan (Pope y Vasquez, 2016).

La dilución de fronteras entre apoyo general y acto clínico degrada la calidad y confunde expectativas. Los marcos de ética profesional insisten en la necesidad de criterios claros para distinguir psicoeducación o bienestar digital de intervención psicológica y advierten que el apoyo automatizado no equivale a práctica clínica regulada. La regla es sobria y conocida. Donde hay evaluación de riesgo, orientación diagnóstica, indicación de técnicas, seguimiento terapéutico o recomendaciones que condicionan decisiones de salud, rigen deberes propios de la psicología profesional y no simples pautas de servicio tecnológico (Knapp, Gottlieb, Handelsman y VandeCreek, 2012).

La ética aplicada añade un punto ineludible. La responsabilidad es personal y no se delega en un algoritmo. Empatía, juicio prudente, obligación de advertencia, manejo de rupturas del vínculo, documentación y supervisión son elementos estructurales del trabajo clínico. Sin un profesional que asuma y responda por la intervención, esos valores quedan incompletos y la rendición de cuentas se diluye en términos de servicio y licencias de uso. Por ello, aun cuando se utilicen herramientas digitales, la conducción clínica debe permanecer bajo responsabilidad humana con competencias demostrables y con protocolos de escalado ante señales de daño o de riesgo (Fisher, 2017).

El desafío para colegios profesionales y reguladores consiste en proteger a la población sin bloquear innovaciones útiles. Ese equilibrio pasa por definir con precisión el alcance permitido de los sistemas de bienestar, exigir validación y supervisión para cualquier función clínica, prohibir la publicidad que sugiera sustitución profesional y establecer rutas claras de responsabilidad para desarrolladores, integradores y prestadores. Solo así la tecnología podrá aportar valor sin habilitar el ejercicio ilegal de la psicología ni erosionar las garantías que amparan a las personas en contextos de vulnerabilidad.

H. CASOS CLÍNICOS ILUSTRATIVOS

Para examinar con precisión los riesgos derivados del uso de inteligencia artificial en salud mental conviene recurrir a estudios de caso hipotéticos. Estos escenarios no describen pacientes reales, pero se apoyan en la metodología cualitativa del estudio de caso, reconocida por su capacidad para integrar múltiples dimensiones de un fenómeno

clínico en su contexto. Su función no es demostrar causalidad, sino explorar de qué modo los sistemas automatizados pueden influir en la dinámica terapéutica, la percepción del malestar y las decisiones de búsqueda de ayuda.

El valor analítico de estos casos reside en que, pese a su carácter ficticio, recogen patrones observados en la práctica profesional y en reportes recientes sobre el uso de aplicaciones de salud digital. Permiten traducir abstracciones éticas y técnicas en situaciones comprensibles, donde las consecuencias de errores de diseño o de uso se vuelven tangibles. Así, cada caso puede entenderse como una herramienta pedagógica que pone a prueba los límites de la intervención automatizada y muestra la diferencia entre acompañamiento tecnológico y tratamiento psicológico propiamente dicho.

A través de esta aproximación narrativa se busca ilustrar tres tipos de riesgo: la dependencia emocional hacia chatbots en adolescentes, la desregulación de pacientes con antecedentes traumáticos frente a respuestas automatizadas inadecuadas y la exposición de información sensible en contextos sin supervisión profesional. Cada uno de estos ejemplos contribuye a visibilizar cómo las limitaciones técnicas y conceptuales de la inteligencia artificial pueden traducirse en daños psicológicos concretos si no se aplican salvaguardas éticas y regulatorias adecuadas.

H.1. CASO DE IDEACIÓN SUICIDA NO DETECTADA

En febrero de 2024 un joven estadounidense de catorce años se quitó la vida tras meses de interacción intensa con un chatbot de inteligencia artificial configurado como personaje de ficción. Los registros de conversación muestran que el adolescente expresó ideación suicida de manera explícita en múltiples ocasiones y que el sistema respondió con frases ambiguas que no activaron protocolos de derivación ni alertas a supervisión humana. La familia presentó demanda contra la empresa desarrolladora alegando diseño negligente y falta de salvaguardas adecuadas para detectar y responder ante crisis (The New York Times, 2024).

Este caso documenta un patrón de falla identificado en revisiones sobre detección de riesgo suicida en plataformas digitales. Los estudios muestran que entre treinta y cuarenta por ciento de expresiones de ideación suicida en aplicaciones de salud mental utilizan lenguaje indirecto, metafórico o eufemístico que elude detección por palabras clave (Bernert et al., 2020). La literatura sobre líneas de crisis subraya que la identificación

temprana se apoya en sensibilidad clínica que integra señales indirectas, cambios de tono, quiebres del relato y oscilaciones afectivas, capacidad que mejora con entrenamiento específico y supervisión pero que excede el alcance de sistemas orientados por correspondencias semánticas literales (Gould et al., 2013).

El análisis del caso revela múltiples puntos de falla. El sistema carecía de mecanismos para detectar patrones de uso problemático como frecuencia excesiva de interacción, aislamiento social progresivo y dependencia emocional hacia el agente artificial. No existían umbrales de escalado automático a intervención humana cuando aparecían menciones explícitas de muerte o suicidio. La arquitectura del producto priorizó inmersión y retención del usuario sobre protocolos de seguridad, un modelo de diseño incompatible con contextos de salud mental donde la vulnerabilidad del usuario exige salvaguardas activas.

La valoración clínica permite sostener la duda cuando el lenguaje es equívoco, ampliar la exploración con preguntas abiertas y ajustar la respuesta al contexto vital. Los marcos sobre señales de alerta insisten en que indicios no literales y evaluación dinámica del riesgo son decisivos (Rudd et al., 2006). La ausencia de mediación profesional ilustra un límite estructural cuando el problema exige lectura profunda del lenguaje, integración de contexto y decisiones de escalado inmediato. La tecnología puede apoyar registrando información y facilitando acceso a recursos, pero no sustituye el juicio clínico que reconoce matices, activa protocolos de seguridad y acompaña de manera responsable durante crisis.

H.2. CASO DE DEPENDENCIA EMOCIONAL DEL SISTEMA

Un hombre de treinta y cinco años con historia de vínculos interpersonales inestables comienza a usar una aplicación de apoyo emocional disponible las veinticuatro horas. Al inicio la herramienta le sirve para atravesar picos de ansiedad antes de reuniones laborales o después de discusiones familiares. Con el correr de los meses el uso se intensifica hasta consultar al sistema varias veces por día ante cualquier malestar menor. La sensación de alivio inmediato resulta tranquilizadora, pero empieza a evitar encuentros con amigos, posterga conversaciones difíciles y reduce actividades que antes le daban sostén. El dispositivo pasa de recurso ocasional a compañía constante.

Este patrón está documentado en estudios sobre uso problemático de aplicaciones de salud mental. Una revisión sistemática sobre adherencia a intervenciones digitales identificó que entre 15 y 20 % de usuarios de aplicaciones de apoyo emocional desarrollan patrones de uso compulsivo caracterizados por aumento progresivo del tiempo conectado, urgencia por consultar y deterioro de funcionamiento social y laboral (Torous et al., 2020). La investigación sobre tecnología persuasiva muestra que diseños orientados a mantener al usuario conectado el mayor tiempo posible mediante notificaciones frecuentes, recompensas intermitentes y disponibilidad ilimitada pueden activar circuitos de dependencia similares a los observados en trastornos de uso de internet y juego problemático (Montag et al., 2019).

El escenario ilustra una paradoja documentada en salud digital. Las métricas de la aplicación pueden mostrar aparente éxito con autorreportes de malestar que disminuyen a corto plazo y frecuencia de uso creciente, mientras el repertorio de afrontamiento se estrecha. La persona depende de confirmación automática para calmarse y pierde práctica en habilidades que requieren demora, exposición graduada y negociación con otros. El alivio repetido refuerza evitación de situaciones que podrían fortalecerse mediante contacto humano real, produciendo un efecto boomerang donde mejora inmediata convive con empeoramiento funcional expresado en aislamiento, menor tolerancia a la frustración y empobrecimiento de vínculos. Este patrón subraya la importancia de monitorizar no solo síntomas sino también indicadores de funcionamiento social y autonomía en usuarios de tecnologías de salud mental.

H.3. CASO DE DIAGNÓSTICO ERRÓNEO POR SESGO ALGORÍTMICO

Una mujer de veintiocho años utiliza una aplicación de inteligencia artificial para evaluar fatiga, irritabilidad y dificultades de concentración. La herramienta, entrenada con datos mayoritariamente urbanos y culturalmente homogéneos, procesa sus respuestas sin atender particularidades de contexto. En su relato aparecen experiencias de discriminación laboral y barreras sociales que incrementan la frustración. El sistema lee esas referencias como hostilidad y las clasifica como rasgos propios de un trastorno de personalidad. Cuando describe apoyarse en prácticas espirituales y comunitarias de su entorno, la aplicación lo codifica como pensamiento irracional y dependencia.

Este patrón refleja hallazgos documentados sobre disparidades en sistemas de inteligencia artificial para salud mental. Un estudio sobre precisión diagnóstica de algoritmos de

aprendizaje automático en poblaciones diversas encontró que la exactitud disminuía entre veinte y treinta y cinco por ciento en minorías étnicas comparado con población blanca, principalmente por subrepresentación en conjuntos de entrenamiento y ausencia de variables contextuales como experiencias de discriminación (Chen et al., 2021). La investigación sobre sesgo en tecnologías de salud muestra que clasificadores entrenados con datos de países de altos ingresos tienden a patologizar expresiones culturales normativas de malestar en contextos no occidentales, generando sobrepredacción de trastornos graves y subdetección de factores de estrés psicosocial (Gianfrancesco et al., 2018).

El sesgo de datos produce una cadena de errores porque la herramienta no reconoce que discriminación social y estructural tiene efectos directos sobre salud mental con manifestaciones clínicas específicas que requieren lectura culturalmente informada. La literatura muestra que exposición sostenida a discriminación se asocia con estrés elevado, síntomas depresivos y ansiedad articulados con identidad y posición social, lo que obliga a evaluación que no reduzca el sufrimiento a rasgos individuales descontextualizados (Williams y Mohammed, 2009). El trauma derivado de experiencias raciales presenta patrones que requieren competencias culturales específicas para no confundir prácticas de sostén comunitario con signos de patología (Carter, 2007).

El caso ilustra cómo modelos que universalizan patrones particulares inducen diagnósticos erróneos y obstaculizan atención adecuada. Una evaluación responsable demanda ampliar anamnesis con preguntas sobre discriminación, redes de apoyo y significados culturales, ajustando interpretación de síntomas a esa trama. La tecnología puede sumar valor si se valida en poblaciones diversas e integra salvaguardas que obliguen a revisar salidas cuando el contexto sugiere factores de riesgo psicosocial, pero no sustituye la lectura clínica que reconoce relación entre contexto, identidad y experiencia de malestar.

I. ALTERNATIVAS RESPONSABLES Y COMPLEMENTARIAS

Reconocer las limitaciones de la inteligencia artificial como sustituto del psicólogo no equivale a rechazar su valor potencial. La clave está en definir con precisión su función dentro del proceso terapéutico y en diseñar marcos de integración que fortalezcan la

práctica profesional sin desplazarla. La tecnología puede ampliar acceso, mejorar continuidad del cuidado y aportar herramientas útiles para evaluación y seguimiento, siempre que su implementación respete los principios éticos, clínicos y jurídicos que rigen la atención psicológica.

Una vía responsable consiste en utilizar la inteligencia artificial como apoyo al trabajo del profesional, no como reemplazo. Los sistemas de análisis de lenguaje pueden ayudar a detectar patrones de cambio emocional a lo largo del tiempo, alertar sobre variaciones bruscas en el discurso o registrar tendencias que el terapeuta evalúa con criterio clínico. Las aplicaciones de autocuidado pueden servir como instrumentos complementarios para reforzar estrategias de afrontamiento entre sesiones, promover hábitos saludables o facilitar acceso a información verificada.

El principio orientador debe ser la colaboración asistida. La tecnología contribuye cuando amplifica la capacidad del profesional para prevenir, intervenir o acompañar, pero pierde legitimidad cuando se presenta como alternativa equivalente al vínculo humano. Los sistemas automatizados deberían operar bajo supervisión y con trazabilidad suficiente para garantizar que las decisiones derivadas de su uso puedan ser revisadas y corregidas.

Un modelo ético de innovación en salud mental requiere participación interdisciplinaria. Psicólogos, psiquiatras, ingenieros, juristas y especialistas en ética aplicada deben intervenir en todas las etapas del diseño y validación de estas herramientas. La inteligencia artificial puede ser un aliado valioso si se la orienta a potenciar la relación terapéutica, no a suprimirla, y si se concibe como medio para humanizar el cuidado mediante tecnología al servicio del bienestar y la dignidad de las personas.

I.1. HERRAMIENTAS DE APOYO PARA PROFESIONALES

La inteligencia artificial aporta valor cuando se integra como soporte del trabajo clínico y no como su reemplazo. Su mayor potencial aparece en tareas complementarias que consumen tiempo y atención del terapeuta, como ordenar documentación, detectar recurrencias temáticas en relatos extensos o seguir la evolución del lenguaje a lo largo de múltiples sesiones. Al delegar estas funciones en sistemas automatizados, el profesional conserva más espacio mental para la escucha, la formulación de casos y la toma de decisiones terapéuticas con sentido clínico (Hsin y cols., 2018).

En los últimos años se ha consolidado la idea de una práctica enriquecida por datos. El juicio profesional sigue siendo el eje de decisión, pero se nutre de información estructurada que las herramientas digitales pueden sintetizar con rapidez. Esta forma de trabajo no sustituye la interpretación del terapeuta, más bien la robustece al ofrecer señales tempranas, panoramas longitudinales y comparaciones intraindividuo que facilitan hipótesis más precisas y revisables (Graham y cols., 2019).

La utilidad de este enfoque ya se observa en los sistemas de retroalimentación en psicoterapia. El monitoreo continuo de indicadores de progreso permite anticipar riesgos de fracaso terapéutico y ajustar el plan antes de que el deterioro se consolide. La incorporación de análisis automatizado puede profundizar esta vigilancia, por ejemplo al señalar patrones sutiles de estancamiento, variaciones atípicas en la alianza o cambios de tono afectivo que pasarían inadvertidos a simple vista. Aun así, el componente decisivo es la mirada clínica, que otorga sentido a los datos, prioriza intervenciones y decide cuándo escalar a medidas de seguridad o cuándo introducir cambios en el encuadre (Bickman, 2008; Lambert, 2010).

I.2. APLICACIONES DE PSICOEDUCACIÓN Y PREVENCIÓN

Una vía prometedora para integrar tecnología en salud mental es la psicoeducación. Los sistemas automatizados pueden ofrecer información clara sobre síntomas frecuentes, pautas de afrontamiento básico y rutas de acceso a recursos comunitarios. En población joven esto resulta especialmente pertinente, ya que el entorno digital se ha convertido en un espacio habitual para buscar orientación y primeros apoyos en temas de bienestar, con tasas de uso sostenidas en servicios en línea cuando la información es comprensible y está bien señalizada hacia ayuda formal cercana al usuario (Burns, Davenport, Durkin, Luscombe y Hickie, 2010).

La evidencia sugiere efectos moderados de las intervenciones en línea para reducir conductas de riesgo y acompañar procesos de cambio cuando operan como complemento y no como sustituto de la atención profesional. Los resultados son más consistentes cuando existe cierta guía, un encuadre de objetivos y una derivación clara a apoyo humano si aparecen señales de deterioro o de riesgo, mientras que los programas sin acompañamiento muestran mayor variabilidad en la adherencia y en el mantenimiento de efectos a mediano plazo (Riper y cols., 2014). Esta lógica ayuda a disminuir barreras de

estigma y culturales, ofrece un primer paso accesible y puede convertir la curiosidad inicial en una consulta oportuna.

El impacto es relevante en contextos universitarios, donde la prevalencia de malestar psicológico supera a la de la población general y los servicios presenciales suelen estar tensionados. Disponer de materiales fiables, autoevaluaciones orientativas y rutas de ayuda locales puede marcar la diferencia entre afrontar en soledad situaciones de vulnerabilidad o acceder a apoyo en tiempo y forma. La tecnología aporta alcance y continuidad, mientras que la supervisión profesional preserva la calidad y evita que herramientas de bajo umbral deriven en reemplazos de la intervención clínica cuando esta es necesaria (Stallman, 2010).

I.3. SISTEMAS DE TRIAJE Y DERIVACIÓN

El triaje digital ofrece una vía concreta para orientar a las personas desde el primer contacto. Algoritmos diseñados con criterios clínicos pueden estimar urgencia, sugerir tipo de atención y priorizar derivaciones según gravedad y riesgo. La experiencia acumulada con instrumentos breves como el Patient Health Questionnaire muestra que los cribados estructurados ayudan a identificar depresión en ámbitos médicos y a iniciar evaluaciones oportunas, lo que respalda el desarrollo de versiones digitales adaptadas para el acceso inicial a servicios de salud mental cuando se acompañan de validación y supervisión adecuadas (Gilbody, Richards, Brealey y Hewitt, 2007).

La masificación de dispositivos móviles amplía el alcance del triaje y permite intervenciones rápidas en poblaciones diversas. Las revisiones indican que los programas entregados por smartphones pueden mejorar detección y compromiso cuando el diseño cuida usabilidad, privacidad y claridad en los siguientes pasos. Además, el registro longitudinal facilita observar cambios y activar alertas tempranas cuando aparecen señales de deterioro, siempre que el profesional conserve la capacidad de revisar y confirmar los hallazgos (Donker, Petrie, Proudfoot, Clarke, Birch y Christensen, 2013).

La eficacia depende de la conexión con redes reales de atención. Un sistema que emite recomendaciones sin asegurar disponibilidad de recursos incrementa frustración y desconfianza. El valor del triaje aparece cuando integra directorios locales actualizados, rutas claras de derivación y protocolos de escalado a contacto humano en riesgo moderado o alto. En esa arquitectura la tecnología no sustituye la psicoterapia, sino que ordena la

demanda, mejora la detección precoz y fortalece la prevención dentro de un rediseño más amplio de la práctica que preserva la centralidad del encuentro clínico y distribuye mejor el esfuerzo profesional (Christensen y Hickie, 2010; Kazdin y Blase, 2011).

I.4. EVIDENCIA FAVORABLE Y POSIBLES BENEFICIOS

La literatura controlada muestra beneficios reales cuando las intervenciones digitales se aplican en contextos bien definidos y con supervisión adecuada. Un metaanálisis en JAMA Psychiatry que revisó cuarenta y cinco ensayos aleatorizados con más de cuatro mil participantes informó tamaños de efecto pequeños a moderados para ansiedad y depresión leves, con mejores resultados en intervenciones guiadas por profesionales que en las completamente automatizadas (Lindhiem et al., 2015). Este patrón sugiere que el acompañamiento humano sigue siendo un componente determinante del cambio clínico.

El ensayo controlado de Woebot con setenta universitarios registró una reducción significativa de síntomas depresivos tras dos semanas de uso, con asignación aleatoria y medidas validadas que respaldan la validez interna del hallazgo (Fitzpatrick et al., 2017). Aun así, la muestra acotada, el seguimiento muy breve y el conflicto de interés declarado limitan su generalización. En una evaluación naturalística de Wysa con más de cien mil usuarios, el sesenta por ciento autoinformó mejoras en bienestar entre las dos y las cuatro semanas de uso en condiciones reales (Inkster et al., 2018). La magnitud muestral favorece la validez externa, pero la ausencia de grupo control, el sesgo de autoselección con tasas de respuesta cercanas al dos por ciento y la falta de información sobre quienes abandonan debilitan de manera sustantiva la fuerza inferencial.

El metaanálisis más amplio sobre chatbots en salud mental sintetizó veintidós ensayos controlados con casi dos mil participantes y halló efectos pequeños frente a controles pasivos, pero diferencias no significativas frente a controles activos o psicoeducación escrita (Linardon et al., 2019). El resultado indica que una porción relevante del beneficio proviene de factores inespecíficos como atención, estructura y expectativas más que de mecanismos propiamente atribuibles a la inteligencia artificial.

La evidencia identifica moderadores consistentes. La supervisión profesional regular incrementa de forma marcada la retención frente al uso autónomo. Los programas de cuatro a ocho semanas rinden mejor que los de uso indefinido. Los efectos se observan en poblaciones con sintomatología leve o moderada y se diluyen en cuadros graves. La

integración con servicios presenciales supera la performance de aplicaciones utilizadas en solitario. Los protocolos psicoeducativos estructurados muestran mejores resultados que la conversación abierta sin guía clínica explícita (Graham et al., 2019).

El sesgo de publicación sigue siendo un problema central. Análisis en salud digital estiman que una proporción considerable de estudios con resultados nulos no llega a publicarse y que los hallazgos positivos suelen concentrarse en equipos con vínculos financieros con desarrolladores, lo que obliga a interpretar con prudencia la magnitud de los efectos reportados (Franco et al., 2014). Muchos ensayos dependen de autorreportes sin evaluación ciega, utilizan seguimientos menores a tres meses y no registran eventos adversos de manera sistemática. Las tasas de abandono son elevadas y rara vez se aplican análisis por intención de tratar, con el consiguiente riesgo de sobreestimar la eficacia.

Tres revisiones independientes coinciden en que la calidad metodológica promedio es baja a moderada, faltan estudios de implementación en sistemas de salud y la evidencia en poblaciones vulnerables es prácticamente inexistente (Bendig et al., 2019; Abd-Alrazaq et al., 2020; Boucher et al., 2021). La heterogeneidad de diseños impide metaanálisis robustos y la ausencia de protocolos estandarizados dificulta comparaciones directas entre plataformas.

Aun con estas limitaciones, los hallazgos justifican explorar integraciones prudentes donde la tecnología actúe como complemento y no como sustituto. El valor aparece cuando se delimitan funciones específicas, se mantiene supervisión profesional continua, se fijan umbrales claros de derivación y se monitoriza el desempeño con métricas clínicas relevantes. La promesa no está en reemplazar la psicoterapia, sino en ampliar el acceso a recursos de primer nivel, facilitar la detección temprana y sostener estrategias de autocuidado entre sesiones dentro de un modelo integrado que preserva la centralidad del vínculo humano.

J. RECOMENDACIONES PARA LA PRÁCTICA PROFESIONAL

La integración de herramientas de inteligencia artificial debe partir de delimitación clara de funciones. Conviene especificar qué tarea realiza la tecnología y qué decisiones permanecen bajo responsabilidad humana, evitando toda presentación que sugiera sustitución del vínculo terapéutico. Esta delimitación se comunica al paciente en lenguaje

comprensible y se revisa cuando cambian las capacidades del sistema o el estado clínico de la persona. El consentimiento se entiende como proceso renovable y revocable, con explicación explícita de alcances, límites y tratamiento de datos sensibles (Beauchamp y Childress, 2019).

La selección de herramientas se apoya en evidencia suficiente para el caso de uso. Antes de su adopción se realiza una prueba piloto acotada y se fijan criterios de éxito clínico que no se confundan con métricas de uso. Los indicadores incluyen evolución sintomática, alianza percibida, eventos adversos y tiempos de derivación, con revisión periódica y documentación de decisiones (Lambert, 2010).

La seguridad exige rutas de escalado bien definidas. Toda interacción automatizada que detecte señales de deterioro activa contacto humano con prioridad temporal adecuada. En población adolescente, trauma complejo o trastornos de personalidad, el uso se restringe a funciones auxiliares bajo supervisión estrecha. En crisis, la herramienta actúa como puente hacia contención profesional inmediata, con registro trazable de alertas.

La justicia y equidad orientan el diseño. Se evalúa desempeño diferencial por subpoblaciones y se corrigen sesgos cuando se detectan brechas. Se garantiza accesibilidad razonable y se evita instaurar circuitos de segunda categoría para quienes no pueden pagar atención presencial (Daniels, 2007).

La gobernanza de datos se rige por minimización, finalidad delimitada y conservación limitada. Se informa quién accede a los registros, con qué propósito y bajo qué base jurídica. La reutilización con fines ajenos al cuidado se somete a controles independientes y auditorías que garanticen trazabilidad (Jobin, Ienca y Vayena, 2019; UNESCO, 2021).

La competencia profesional incluye alfabetización tecnológica. El equipo clínico recibe formación específica sobre alcances y límites de los sistemas utilizados, riesgos previsibles y comunicación honesta con los pacientes. La implementación se concibe como proyecto interdisciplinario con responsabilidades definidas y revisión periódica de desempeño y seguridad. La tecnología aporta valor cuando amplifica la capacidad del profesional para cuidar y no cuando la reemplaza.

J.1. EVALUACIÓN CRÍTICA DE HERRAMIENTAS TECNOLÓGICAS

Antes de incorporar una aplicación clínica conviene examinar su fundamento empírico y sus límites. Las guías profesionales recomiendan valorar qué problema resuelve, con qué población se probó, qué resultados clínicos midió y si existen evaluaciones independientes que reproduzcan sus hallazgos. También piden revisar la transparencia del algoritmo, la procedencia y calidad de los datos de entrenamiento, la trazabilidad de versiones y la presencia de reportes sobre efectos adversos y eventos de seguridad. La decisión responsable se apoya en documentación verificable y no en descripciones comerciales (American Psychological Association, 2017).

Además, importa la adecuación cultural y demográfica. Una herramienta validada en muestras homogéneas puede distorsionar síntomas cuando se usa en contextos distintos, especialmente si no reporta desempeño por subgrupos ni estrategias de mitigación de sesgos. Un examen crítico incluye analizar variables de exclusión, métricas empleadas, tasa de falsos negativos en escenarios de riesgo y capacidad de interoperar con la historia clínica para que el profesional interprete los datos con contexto suficiente (Reed, McLaughlin y Milholland, 2000).

Diversos informes advierten sobre promesas sobredimensionadas y controles insuficientes. La adopción apresurada de sistemas opacos puede amplificar desigualdades y desplazar la atención desde resultados clínicos hacia métricas de uso. Por eso conviene exigir publicación de métodos, auditorías externas y conflicto de intereses declarado, así como planes de mejora continua que contemplen supervisión humana y retiro del sistema si aparecen fallos graves (Crawford y cols., 2016). La innovación valiosa se distingue por su rendimiento reproducible y por su integración segura en los circuitos asistenciales, no por la novedad técnica aislada. En este sentido, una lectura clínica de la evidencia y una gobernanza sólida pesan más que el marketing de alto impacto (Topol, 2019).

La dimensión legal y regulatoria completa la evaluación. Persisten vacíos normativos y marcos en transición que obligan a prudencia. Antes de usar una herramienta conviene confirmar su estatus regulatorio, el encuadre de protección de datos, los términos de responsabilidad frente a daños y las rutas de escalado a atención humana. Hasta que estos puntos estén bien resueltos, la adopción debe ser gradual, con pilotos acotados, criterios de éxito clínico definidos a priori y mecanismos de salida claros.

J.2. COMUNICACIÓN CON PACIENTES SOBRE TECNOLOGÍA

El uso de aplicaciones de autoayuda debe integrarse en un diálogo explícito entre profesional y paciente. Es responsabilidad del psicólogo explicar de manera comprensible que ofrecen los recursos de psicoeducación y autocuidado y en qué se diferencian de intervenciones terapéuticas que requieren evaluación, formulación de caso y seguimiento clínico. Este encuadre evita confusiones, alinea expectativas y protege la alianza al dejar claro que la tecnología puede acompañar, pero no sustituye la relación y el juicio profesional que sostienen el proceso de cambio (Maheu, Pulier, Wilhelm, McMenamin y Brown-Connolly, 2004).

La experiencia acumulada en la expansión de la telepsicología durante la pandemia mostró oportunidades y límites. Diversos servicios a distancia demostraron eficacia comparable a formatos presenciales cuando se organizaron con criterios clínicos claros, confidencialidad adecuada y rutas de escalado. Sin embargo, también se evidenciaron riesgos asociados a la sobreconfianza en herramientas no validadas y a la falta de protocolos para crisis. Comunicar estos matices es parte del consentimiento informado continuo y ayuda a que el paciente comprenda qué puede esperar de una aplicación y qué situaciones requieren contacto humano directo y oportuno (Pierce, Perrin, Tyler, McKee y Watson, 2021).

Conviene acordar políticas de uso paralelo desde el inicio del tratamiento. Se explicita qué tipo de aplicaciones se recomienda emplear, con qué propósito y en qué momentos, y se desaconseja el uso de sistemas que prometen diagnóstico o tratamiento sin supervisión. Se fija un plan de seguridad para crisis que prioriza contacto humano, se aclara cómo se compartirán registros relevantes en sesión y se advierte que métricas de uso o de ánimo no reemplazan la conversación clínica. También se incluyen pautas de privacidad y límites de comunicación entre sesiones para evitar derivar la regulación emocional cotidiana al intercambio con la herramienta.

Este encuadre convierte a la tecnología en aliada. El paciente dispone de recursos de bajo umbral para psicoeducación y hábitos saludables, mientras la terapia conserva su núcleo relacional y su foco en significados, ambivalencias y decisiones. La claridad en la comunicación reduce expectativas irreales, previene sustituciones de bajo valor y sostiene la responsabilidad profesional sobre las intervenciones que impactan en la salud mental.

J.3. COLABORACIÓN INTERDISCIPLINARIA

El desarrollo responsable de tecnología en salud mental requiere equipos mixtos desde el inicio. La complejidad del sufrimiento humano no se traduce de forma directa en requisitos técnicos, por eso resulta imprescindible que psicólogos y otros profesionales clínicos participen en la definición del problema, en la priorización de funciones y en la validación de efectos. Su intervención ayuda a que la lógica del mercado y la búsqueda de eficiencia no desplacen el foco en el bienestar del paciente y en la seguridad del proceso terapéutico. La literatura sobre diseño centrado en la salud mental insiste en integrar co-diseño con usuarios, pruebas de campo y ciclos de evaluación que contemplen contexto clínico real y no solo entornos de laboratorio, con especial atención a riesgo, tono relacional y carga emocional de las interacciones digitales (Doherty, Coyle y Matthews, 2011).

La colaboración no se limita a correcciones de lenguaje o a ajustes de interfaz. Implica traducir modelos clínicos en decisiones de producto y, al mismo tiempo, adaptar los flujos de trabajo para que la herramienta encuentre su lugar en la atención. Cuando ingenieros, especialistas en ética y profesionales de la salud trabajan en conjunto, aparecen soluciones más realistas, con mayor aceptabilidad y mejores probabilidades de adopción sostenida. Las tecnologías de intervención conductual muestran un potencial notable para ampliar el alcance de la psicología, siempre que se diseñen para necesidades reales, con metas clínicas explícitas, mecanismos de retroalimentación y rutas claras de escalado a atención humana en caso de deterioro o riesgo emergente (Schueller, Muñoz y Mohr, 2013).

Este enfoque converge con el modelo de intervención digital, que propone un marco integrado donde los recursos tecnológicos se combinan con principios clínicos. La herramienta se concibe como un medio para apoyar evaluación, prevención y seguimiento, mientras la formulación de caso, la priorización terapéutica y las decisiones de riesgo permanecen bajo responsabilidad profesional. El modelo destaca la importancia de alinear funcionalidades con objetivos clínicos, definir indicadores de éxito relevantes para el paciente y asegurar trazabilidad para auditar desempeño y seguridad a lo largo del tiempo. De este modo, la innovación no se separa de la evidencia ni de la ética, y la tecnología refuerza la relación terapéutica en lugar de sustituirla (Mohr, Schueller, Montague, Burns y Rashidi, 2014).

K. PROPUESTAS PARA UNA REGULACIÓN RESPONSABLE

Proteger el bienestar público frente a la expansión de la inteligencia artificial en salud mental exige un marco normativo capaz de equilibrar innovación y seguridad. La ausencia de reglas claras expone a los usuarios a riesgos clínicos y éticos, mientras que regulación excesivamente restrictiva podría sofocar avances con potencial social positivo. La cuestión es diseñar estructuras que garanticen transparencia, rendición de cuentas y control profesional sin frenar el desarrollo tecnológico.

Un enfoque regulatorio responsable debe combinar tres dimensiones. La ética define límites sobre lo que puede automatizarse y bajo qué condiciones. La clínica establece criterios mínimos de validez, eficacia y supervisión profesional. La técnica regula trazabilidad de algoritmos, gestión de datos sensibles y obligación de auditorías independientes. Estas dimensiones deben interactuar priorizando seguridad del paciente e integridad de la práctica profesional.

Un punto clave consiste en adoptar regulación por niveles que distinga entre tecnologías de bajo, medio y alto riesgo según potencial de impacto en derechos fundamentales. Las aplicaciones informativas podrían operar con requisitos de registro y transparencia, mientras aquellas que simulan o reemplazan intervenciones clínicas deberían someterse a evaluaciones previas de seguridad, certificaciones equivalentes a dispositivos médicos y mecanismos de monitoreo continuo.

Resulta indispensable definir responsabilidades compartidas entre desarrolladores, proveedores, autoridades sanitarias y colegios profesionales. Cada actor debe asumir obligaciones precisas en diseño seguro, gestión de datos, supervisión, respuesta ante fallos, certificación y formación continua. La regulación debe prever sanciones proporcionales ante daños, omisiones o prácticas engañosas.

Cualquier marco legal futuro deberá incorporar mecanismos de participación ciudadana y consulta interdisciplinaria. Solo un diálogo sostenido entre ciencia, tecnología y derechos humanos puede asegurar que la innovación sirva al bienestar. Una regulación verdaderamente responsable construye confianza social mediante transparencia, supervisión y respeto por la vulnerabilidad humana.

K.1. CATEGORIZACIÓN DE SERVICIOS DIGITALES DE SALUD MENTAL

La regulación de la inteligencia artificial en salud mental necesita categorías claras según alcance y nivel de riesgo. No es razonable tratar del mismo modo a una aplicación que ofrece pautas básicas de relajación y a un sistema que declara capacidad para diagnosticar o recomendar tratamientos. Un enfoque proporcional, orientado a la confianza y a la protección de derechos, permite graduar exigencias en función del impacto esperado sobre las personas y del contexto de uso, tal como proponen las directrices europeas para una inteligencia artificial confiable (High-Level Expert Group on Artificial Intelligence, 2019).

En el extremo de menor riesgo se ubican las herramientas de carácter psicoeducativo y de promoción de autocuidado. Su régimen puede asimilarse al de productos de bienestar, siempre que cumplan condiciones mínimas de transparencia. Deben declarar con claridad su naturaleza informativa, describir límites de uso, evitar cualquier insinuación de equivalencia terapéutica y advertir rutas de derivación cuando el malestar excede el alcance de la herramienta. La evaluación clínica no es exigible en el mismo nivel que para un producto sanitario, pero sí lo es la veracidad de las afirmaciones y la ausencia de prácticas de marketing que induzcan a error sobre su finalidad (U.S. Food and Drug Administration, 2019).

En el polo de mayor riesgo se encuentran los sistemas que afirman diagnosticar trastornos o proveer intervenciones terapéuticas específicas. En estos casos corresponde un control robusto: evidencia científica sólida sobre eficacia y seguridad en poblaciones pertinentes, validaciones independientes, gestión de riesgos centrada en seguridad del paciente, supervisión profesional efectiva y vigilancia poscomercialización con reporte de eventos adversos. Sin estos mecanismos, la exposición del usuario a daños clínicos previsibles —diagnósticos erróneos, retraso en la búsqueda de ayuda, recomendaciones inadecuadas— resulta incompatible con un marco ético y jurídico de protección en salud mental, como advierte la Organización Mundial de la Salud para el ámbito de la inteligencia artificial aplicada a salud (World Health Organization, 2021).

La presentación pública es parte integral de la categorización. No basta con revisar funcionalidades técnicas. Importa también cómo el producto se describe y qué expectativas crea. El Foro Internacional de Reguladores de Dispositivos Médicos ha propuesto evaluar el riesgo considerando uso previsto, contexto y consecuencias de una interpretación errónea. Así, una aplicación que se publicita como sustituto de la psicoterapia profesional, aunque en la práctica ofrezca mensajes genéricos de autoayuda,

debe someterse a estándares equiparables a los de alto riesgo, porque la confusión inducida puede tener efectos graves para la salud (International Medical Device Regulators Forum, 2021).

Esta gradación no solo organiza el ecosistema digital. También construye confianza pública. Un sistema responsable garantiza que las personas entienden con qué tipo de herramienta interactúan, cuáles son sus límites y qué nivel de protección regulatoria lo respalda. La proporcionalidad regulatoria, aplicada con rigor técnico y sensibilidad clínica, es la vía para habilitar innovaciones útiles sin sacrificar la seguridad ni la calidad del cuidado.

K.2. REQUISITOS DE TRANSPARENCIA Y CONSENTIMIENTO INFORMADO

La transparencia es un eje regulatorio ineludible cuando la interacción se media por sistemas automatizados. Las directrices europeas sobre decisiones automatizadas exigen informar de manera clara y accesible que el usuario está frente a un sistema no humano, detallar su alcance real y advertir sus limitaciones en comparación con la intervención clínica, con el fin de reducir confusiones y evitar equivalencias engañosas con la psicoterapia profesional (Article 29 Data Protection Working Party, 2018). Esta obligación no se satisface con fórmulas genéricas. Requiere lenguaje comprensible, ejemplos situados y señalización explícita de los supuestos de uso responsable.

El debate sobre un derecho a explicación ha ganado visibilidad. La idea es ofrecer a las personas una comprensión suficiente de cómo y por qué un algoritmo produce determinados resultados. La literatura reconoce su valor normativo, pero también sus límites prácticos, ya que muchos modelos de aprendizaje automático no admiten una explicación simple ni plenamente inteligible para el público no experto. El reto es traducir razonamientos probabilísticos complejos en justificaciones auditables y útiles para tomar decisiones informadas, sin prometer una transparencia técnica que hoy no siempre es posible alcanzar de modo completo (Goodman y Flaxman, 2017; Wachter, Mittelstadt y Floridi, 2017).

El Parlamento Europeo, a través del Artificial Intelligence Act, avanzó en obligaciones específicas de información para sistemas de alto riesgo. El énfasis está en pasar de advertencias abstractas a mecanismos efectivos de comprensión que describan el tipo de servicio, los límites técnicos, los supuestos de funcionamiento y los umbrales que exigen

recurrir a ayuda profesional. Se trata de garantizar que las personas entiendan qué puede y qué no puede hacer la herramienta, y qué vías de atención humana están disponibles cuando la situación lo demande (European Parliament, 2021).

En salud mental estas exigencias se intensifican. Debe indicarse de forma explícita que la inteligencia artificial no sustituye supervisión clínica, que puede fallar en la detección de crisis y que, en ocasiones, opera bajo lógicas comerciales orientadas a retención antes que a beneficio terapéutico. El consentimiento informado no se reduce a aceptar términos extensos. Es un proceso continuo que incluye explicaciones claras, ejemplos prácticos, advertencias sobre riesgos y rutas de derivación. En poblaciones vulnerables como menores, personas con dificultades cognitivas o usuarios en crisis emocional, el consentimiento requiere formatos adaptados y verificación de comprensión, pues de lo contrario quienes más protección necesitan quedan expuestos a diagnósticos inadecuados, confusiones o demoras en la búsqueda de ayuda.

K.3. SUPERVISIÓN PROFESIONAL OBLIGATORIA

Los sistemas que prometen intervenciones terapéuticas no deben operar de manera autónoma. La posición de referencia sostiene que toda aplicación clínica basada en inteligencia artificial requiere supervisión humana directa para preservar la responsabilidad sobre cada decisión y para sostener el estándar de cuidado propio de la práctica sanitaria. La Asociación Médica Mundial subraya que la inteligencia artificial solo puede actuar como apoyo y que la conducción clínica permanece en manos del profesional responsable, con capacidad de validar recomendaciones, corregir desvíos y decidir el curso de acción ante incertidumbre o riesgo elevado (World Medical Association, 2018).

Esta exigencia se complementa con obligaciones para desarrolladores y proveedores. El Código de Ética de la Association for Computing Machinery recuerda que el diseño y el despliegue de sistemas tecnológicos deben priorizar la seguridad y el bienestar de las personas por encima de beneficios operativos o comerciales. La funcionalidad técnica por sí sola no legitima un uso clínico. Hace falta definir integración en el circuito asistencial, asignación de responsabilidades, criterios de corte para derivación y mecanismos de corrección cuando aparecen fallos o efectos no deseados en poblaciones reales (Association for Computing Machinery, 2018).

Los estándares de la IEEE avanzan en un punto crucial. La supervisión no es un trámite, es un proceso continuo que exige acceso suficiente a información técnica para evaluar idoneidad, límites y condiciones de uso. El profesional que supervisa debe conocer qué resultados produce el sistema y con qué criterios los genera. Sin transparencia sobre datos, versionado y comportamiento esperado, la supervisión se vuelve nominal y pierde eficacia para proteger a los pacientes y para aprender de los incidentes que inevitablemente ocurren al operar en contextos complejos de salud mental (Institute of Electrical and Electronics Engineers, 2020).

La seguridad del paciente exige además rutas claras de transición a atención humana inmediata. La Asociación Médica Americana insiste en que, ante señales de ideación suicida, irrupción de síntomas psicóticos o deterioro funcional grave, el sistema debe activar protocolos que conecten sin demora con profesionales disponibles y que documenten la cadena de acciones para garantizar continuidad de cuidado y trazabilidad. Dejar a una persona en crisis atrapada en respuestas automatizadas socava la confianza y expone a daños previsibles que la regulación y la ética profesional buscan evitar de manera explícita (American Medical Association, 2021).

El equilibrio entre innovación y seguridad es alcanzable cuando la supervisión es real y operativa. La tecnología puede ampliar acceso, ordenar demanda y reforzar prevención, pero no sustituye la responsabilidad clínica que solo pueden ejercer profesionales acreditados. Un marco de supervisión continua, con transparencia técnica y protocolos de escalados efectivos, permite integrar estas herramientas de modo responsable sin erosionar la confianza que sostiene el trabajo terapéutico.

L. REFLEXIONES Y DIRECCIONES FUTURAS

El análisis realizado muestra que la inteligencia artificial, en su estado actual, no puede ni debe figurar como sustituto de la práctica psicológica. La diferencia entre traducir regularidades estadísticas y comprender el sentido del sufrimiento es un límite ontológico que separa correlación y significado. Mientras la psicoterapia trabaja con historia, vínculo y contexto, los modelos operan con aproximaciones probabilísticas que sin conducción clínica tienden a simplificar lo que exige delicadeza interpretativa. La confianza social en

estas herramientas se sostiene en gobernanza ética robusta con responsabilidades claras y mecanismos verificables de seguridad.

El futuro razonable no niega la tecnología, la encuadra. Resulta deseable avanzar hacia integraciones que refuercen prevención, seguimiento y continuidad del cuidado sin erosionar la centralidad del encuentro humano. Esto requiere investigación aplicada con diseños longitudinales y resultados clínicamente relevantes, auditorías independientes que examinen desempeño por subpoblaciones y documenten eventos adversos, con publicación abierta de métodos y fallos. La innovación será creíble cuando incorpore co-diseño con pacientes y clínicos, someta a prueba sus límites en escenarios reales y acepte retirar funcionalidades que muestren daño o valor marginal.

La prioridad está en sistemas que mejoren detección temprana, organicen información para juicio clínico y ofrezcan apoyos entre sesiones sin sustituir deliberación compartida. Ello implica supervisión profesional continua, protocolos de escalado en crisis y minimización estricta de datos sensibles. La ética de implementación debe traducirse en decisiones concretas sobre qué hace la herramienta, cuándo debe ceder el paso a intervención humana y qué derechos asisten al usuario en materia de privacidad. La regulación debe acompañar con criterios de proporcionalidad por riesgo, trazabilidad técnica y sanciones efectivas cuando la publicidad induzca a confusión o cuando el diseño priorice retención sobre bienestar.

La dirección es nítida. Menos retórica y más evidencia, menos sustitución y más colaboración, menos opacidad y más rendición de cuentas. Si el campo sostiene estos principios, la inteligencia artificial podrá convertirse en aliado útil para ampliar acceso y mejorar continuidad, mientras la psicoterapia conserva su núcleo insustituible de presencia, escucha y responsabilidad compartida.

L.1. SÍNTESIS DE HALLAZGOS PRINCIPALES

El examen del campo muestra límites decisivos que impiden considerar a la inteligencia artificial como sustituto de la psicoterapia. En el plano técnico, los sistemas actuales operan sobre correlaciones y carecen de experiencia encarnada. Pueden detectar palabras frecuentes, pero no acceden al significado implícito ni distinguen ironía, ambivalencias sostenidas por el silencio o metáforas enraizadas en trasfondo cultural específico. Esta brecha no se corrige con más datos porque está anclada en ausencia de subjetividad y de

participación en el mundo social que otorga sentido a los signos clínicos (Russell, Dewey y Tegmark, 2015).

En el plano ético, el modelo de negocio dominante prioriza retención y tiempo de uso. La validación constante ofrece alivio rápido, pero desalienta pausa reflexiva y refuerza dependencia, en dirección contraria a un cuidado que busca autonomía y cambios duraderos. Cuando las métricas de éxito se confunden con frecuencia de interacción, el diseño se aleja de objetivos clínicos y aumenta el riesgo de sustituciones de bajo valor.

En el plano jurídico y regulatorio persiste zona gris que expone a los usuarios. No está resuelto si muchas aplicaciones deben regirse como productos de consumo, software sanitario o herramientas de bienestar, y esa indefinición permite que se ofrezcan como alternativas terapéuticas sin evidencia clínica equivalente, sin supervisión profesional efectiva y sin trazabilidad suficiente. La falta de categorías claras y obligaciones proporcionales al riesgo erosiona la confianza (European Commission, 2020).

L.2. IMPLICANCIAS PARA LA POLÍTICA PÚBLICA

Los hallazgos tienen consecuencias directas para el diseño de política sanitaria y regulación de tecnologías emergentes. La gobernanza internacional de la inteligencia artificial sigue siendo desigual y fragmentaria, con marcos que avanzan a ritmos dispares y escasa coordinación transfronteriza. Esa heterogeneidad crea corredores de baja exigencia donde circulan herramientas sin validación suficiente y desplazan la carga del riesgo hacia los usuarios más expuestos (Butcher y Beridze, 2019). La prioridad pública debe situarse en asegurar oferta clínica humana suficiente y accesible. La tecnología puede ampliar alcance y ordenar demanda, pero no sustituye la relación terapéutica ni el juicio profesional.

Las políticas de financiación conviene orientarlas a reforzar redes comunitarias, primer nivel de atención y dispositivos especializados, reservando a las soluciones automatizadas un rol complementario y acotado. Aceptar que poblaciones con menos recursos reciban solo apoyo digital naturaliza un circuito de menor valor y consolida modelo dual donde la calidad queda determinada por capacidad de pago. Evitar ese desenlace exige diseñar salvaguardas distributivas que impidan que la innovación se traduzca en nuevas capas de desigualdad (World Economic Forum, 2020).

El fortalecimiento educativo es otro eje estratégico. La alfabetización crítica sobre tecnologías de salud mental permite comprender alcances y límites, mejora toma de decisiones y reduce dependencia acrítica de herramientas que pueden generar falsas expectativas o retrasar búsquedas de ayuda. Este esfuerzo debe comenzar en ámbitos escolares y extenderse a campañas públicas alineadas con agendas de desarrollo sostenible (OECD, 2019).

La cooperación internacional resulta indispensable para evitar consolidación de estándares desiguales. Avanzar hacia principios compartidos en transparencia, supervisión profesional y evaluación de seguridad puede armonizar expectativas y reducir arbitrajes regulatorios. La elaboración de estándares éticos con participación de expertos clínicos y sociedad civil contribuye a orientar el desarrollo técnico hacia fines compatibles con el bienestar de usuarios de servicios de salud mental (Winfield, 2019).

L.3. AGENDA DE INVESTIGACIÓN FUTURA

Las prioridades de investigación se orientan a cerrar brechas que hoy impiden una integración segura y con sentido clínico. Se necesitan estudios longitudinales que describan efectos sostenidos del uso continuado de aplicaciones basadas en inteligencia artificial. Hace falta medir dependencia tecnológica, cambios en habilidades de afrontamiento, adherencia a tratamientos presenciales y tiempos de acceso a ayuda profesional cuando aparecen señales de deterioro. Estos diseños deben incluir muestras amplias, resultados clínicos relevantes y reportes de eventos adversos (Nature Portfolio, 2018).

Resulta clave avanzar en transparencia operativa y capacidad explicativa útil para la clínica. La confianza en contextos de salud depende de que los profesionales comprendan qué variables influyen en las salidas del sistema, con qué certeza y en qué condiciones conviene escalar a evaluación humana. La investigación sobre explicabilidad en medicina ofrece un marco prometedor para traducir razonamientos algorítmicos en justificaciones auditables y diseñar interfaces que presenten información sin sobrecargar al clínico ni inducir falsas seguridades (Holzinger, Langs, Denk, Zatloukal y Müller, 2019).

Un tercer frente ineludible es el transcultural. La expresión del malestar psicológico varía según contextos y biografías, por lo que modelos entrenados con datos homogéneos tienden a fallar en detección e interpretación. Se requieren cohortes diversas, validaciones

por subpoblaciones y procedimientos sistemáticos para identificar y mitigar sesgos que reproducen desigualdades. Una agenda robusta debe integrar evaluación diferencial de desempeño, co-diseño con comunidades y revisión continua de impacto (Singh, Drummond y Ahmadi, 2022).

Esta agenda se completa con prácticas de ciencia abierta y diseños rigurosos. Ensayos preinscritos, métricas clínicamente significativas, comparadores adecuados, seguimiento a medio y largo plazo y publicación de resultados negativos fortalecen la credibilidad del campo. La coordinación entre investigadores, clínicos y desarrolladores permitirá transformar promesas en conocimiento operativo y decidir dónde la tecnología suma y dónde debe ceder el protagonismo a la relación terapéutica.

M. CONCLUSIÓN

El entusiasmo por la inteligencia artificial nace de su promesa de respuestas ágiles ante problemas complejos. Sin embargo, la experiencia clínica y jurídica muestra que el alivio del sufrimiento psíquico no procede de algoritmos, sino de vínculos humanos capaces de escuchar, contener y otorgar sentido. La ética de la responsabilidad exige prudencia, sobre todo cuando la intervención recae en personas vulnerables. En psicología, la relación terapéutica no es un accesorio sino el núcleo del cambio. Ninguna simulación reproduce su densidad afectiva ni el valor transformador de la confianza mutua.

Los avances informáticos han ampliado la capacidad de procesar información, reconocer patrones y responder en lenguaje natural. Esa potencia entraña un riesgo: confundir coherencia formal con comprensión del sentido. En el plano jurídico, la confusión se traduce en vacíos de responsabilidad, tensiones en protección de datos y dudas sobre el alcance del ejercicio profesional. La falta de marcos claros habilita sistemas que imitan prácticas clínicas sin cumplir estándares equivalentes y deja a los usuarios expuestos a daños sin rutas efectivas de reparación.

El futuro de la salud mental no se juega en una dicotomía entre tradición y tecnología. Requiere integración cuidadosa, donde psicología, derecho e informática coordinen para garantizar prácticas seguras y éticas. La tecnología puede ampliar acceso, fortalecer prevención y apoyar el seguimiento, pero no debe sustituir lo esencial: la presencia humana que reconoce y acompaña el dolor subjetivo. Este horizonte reclama virtudes

prácticas como prudencia, justicia y responsabilidad, ahora aplicadas a entornos atravesados por datos masivos, aprendizaje automático y regulaciones en construcción, con una orientación explícita al florecimiento humano y no solo a la eficiencia técnica (Vallor, 2016). Para los clínicos, esto implica sostener la centralidad de la relación. Para los desarrolladores, diseñar con prioridad en el bienestar. Para los legisladores, establecer normas proporcionales al riesgo que protejan sin sofocar la innovación.

Este artículo no presenta evidencia empírica original ni una revisión sistemática. Propone una reflexión crítica interdisciplinaria que delimita el alcance real de la inteligencia artificial en salud mental, reconoce sus límites y sugiere criterios éticos y jurídicos para integrar herramientas digitales sin desvirtuar lo más valioso de la práctica clínica: el encuentro entre personas.

N. BIBLIOGRAFÍA

Abd-Alrazaq, A., Reddel, M., Alhuwail, D., Alajlani, M., Aziz, S., Ahmed, A., & Sheikh, J. (2020). Effectiveness of chatbots for mental health: Systematic review. Journal of Medical Internet Research, 22(7), e16021.

Abbott, R. (2020). The reasonable robot: Artificial intelligence and the law. Cambridge University Press.

Alter, A. (2017). Irresistible: The rise of addictive technology and the business of keeping us hooked. Penguin Press.

American Medical Association. (2021). AMA principles for augmented intelligence development, deployment and use. AMA House of Delegates.

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct. American Psychological Association.

American Psychological Association. (2017). Guidelines for the practice of telepsychology. American Psychologist, 68(9), 791–800.

Article 29 Data Protection Working Party. (2018). Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. European Commission.

Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). Oxford University Press.

Bendig, E., Erb, B., Schulze-Thising, L., & Baumeister, H. (2019). The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health. Verhaltenstherapie, 29(4), 266–280.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185–5198).

Bickman, L. (2008). A measurement feedback system (MFS) is necessary to improve mental health outcomes. Journal of the American Academy of Child & Adolescent Psychiatry, 47(10), 1114–1119.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, Research & Practice, 16(3), 252–260.

Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. Expert Review of Medical Devices, 18(sup1), 37–49.

Bowlby, J. (1988). A secure base: Parent–child attachment and healthy human development. Basic Books.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15. (Conference on Fairness, Accountability, and Transparency).

Butcher, J., & Beridze, I. (2019). What is the state of artificial intelligence governance globally? The RUSI Journal, 164(5–6), 88–96.

Carter, R. T. (2007). Racism and psychological and emotional injury: Recognizing and assessing race-based traumatic stress. The Counseling Psychologist, 35(1), 13–105.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science, 4*, 123–144.

Christensen, H., & Hickie, I. B. (2010). Using e-health applications to deliver new mental health services. Medical Journal of Australia, 192(11), S53–S56.

Crawford, K., Whittaker, M., Elish, M. C., Barocas, S., Plasek, A., & Ferryman, K. (2016). The AI Now Report: The social and economic implications of artificial intelligence technologies in the near-term. AI Now Institute.

Daniels, N. (2007). Just health: Meeting health needs fairly. Cambridge University Press.

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M. R., & Christensen, H. (2013). Smartphones for smarter delivery of mental health programs: A systematic review. Journal of Medical Internet Research, 15(11), e247.

Doherty, G., Coyle, D., & Matthews, M. (2011). Design and evaluation guidelines for mental health technologies. Interacting with Computers, 23(3), 243–252.

Ekman, P. (2001). Telling lies: Clues to deceit in the marketplace, politics, and marriage (3rd ed.). W. W. Norton.

Erikson, E. H. (1968). Identity: Youth and crisis. W. W. Norton.

European Commission. (2020). White paper on artificial intelligence: A European approach to excellence and trust (COM(2020) 65 final).

European Parliament. (2021). Artificial Intelligence Act: First regulatory framework on artificial intelligence. European Union.

Faden, R. R., & Beauchamp, T. L. (1986). A history and theory of informed consent. Oxford University Press.

Fisher, C. B. (2017). Decoding the ethics code: A practical guide for psychologists (4th ed.). Sage Publications.

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of Medical Internet Research, 21(5), e13216.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19.

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A, 374(2083), 20160360.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. Science, 345(6203), 1502–1505.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. JMIR Mental Health, 5(4), e64.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine, 178*(11), 1544–1547.

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. Journal of General Internal Medicine, 22(11), 1596–1602.

Goodman, K. E., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50–57.

Gould, M. S., Cross, W., Pisani, A. R., Munfakh, J. L., & Kleinman, M. (2013). Impact of applied suicide intervention skills training on the National Suicide Prevention Lifeline. Suicide and Life-Threatening Behavior, 43(6), 676–691.

Gould, M. S., Greenberg, T., Velting, D. M., & Shaffer, D. (2003). Youth suicide risk and preventive interventions: A review of the past 10 years. Journal of the American Academy of Child & Adolescent Psychiatry, 42(4), 386–405.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and dementia care—A national research and development agenda. JAMA Psychiatry, 76(4), 365–366.

Griffiths, K. M., Calear, A. L., & Banfield, M. (2009). Systematic review on Internet Support Groups (ISGs) and depression (1): Do ISGs reduce depressive symptoms? Journal of Medical Internet Research, 11(3), e40.

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.

Herman, J. L. (1992). Trauma and recovery: The aftermath of violence—from domestic abuse to political terror. Basic Books.

High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

Hsin, H., Fromer, M., Peterson, B., Walter, R., Fleck, M., Campbell, A., … & Patel, V. (2018). Transforming psychiatry into data-driven medicine with digital measurement tools. NPJ Digital Medicine, 1(1), 37.

Institute of Electrical and Electronics Engineers. (2020). IEEE standards for artificial intelligence systems. IEEE Computer Society.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation. JMIR mHealth and uHealth, 6(11), e12106.

International Medical Device Regulators Forum. (2021). Software as a Medical Device: Possible framework for risk categorization and corresponding considerations. IMDRF.

Joiner, T. E. (2005). Why people die by suicide. Harvard University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspectives on Psychological Science, 6(1), 21–37.

Kernberg, O. F. (1975). Borderline conditions and pathological narcissism. Jason Aronson.

Kleinman, A. (1988). Rethinking psychiatry: From cultural category to personal experience. Free Press.

Knapp, S. J., Gottlieb, M. C., Handelsman, M. M., & VandeCreek, L. D. (2012). APA handbook of ethics in psychology. American Psychological Association.

Kohut, H. (1971). The analysis of the self: A systematic approach to the psychoanalytic treatment of narcissistic personality disorders. International Universities Press.

Lambert, M. J. (2010). Prevention of treatment failure: The use of measuring, monitoring, and feedback in clinical practice. American Psychological Association.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. Psychotherapy: Theory, Research, Practice, Training, 38(4), 357–361.

Levi, S. (2018). AI liability: Toward a working typology. Georgetown Journal of International Law, 49(4), 1109–1159.

Lindhiem, O., Bennett, C. B., Rosen, D., & Silk, J. (2015). Mobile technology boosts the effectiveness of psychotherapy and behavioral interventions: A meta-analysis. Behavior Modification, 39(6), 785–804.

Linehan, M. M. (1993). Cognitive-behavioral treatment of borderline personality disorder. Guilford Press.

Maheu, M. M., Pulier, M. L., Wilhelm, F. H., McMenamin, J. P., & Brown-Connolly, N. E. (2004). The mental health professional and the new technologies: A handbook for practice today. Lawrence Erlbaum Associates.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.

Marmot, M. (2005). Social determinants of health inequalities. The Lancet, 365(9464), 1099–1104.

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2019). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Internal Medicine, 176(5), 619–625.

Mohr, D. C., Burns, M. N., Schueller, S. M., Clarke, G., & Klinkman, M. (2013). Behavioral intervention technologies: Evidence review and recommendations for future research in mental health. General Hospital Psychiatry, 35(4), 332–338.

Mohr, D. C., Schueller, S. M., Montague, E., Burns, M. N., & Rashidi, P. (2014). The behavioral intervention technology model: An integrated conceptual and technological

framework for eHealth and mHealth interventions. Journal of Medical Internet Research, 16(6), e39.

Montag, C., Lachmann, B., Herrlich, M., & Zweig, K. (2019). Addictive features of social media/messenger platforms and freemium games against the background of psychological and economic theories. *International Journal of Environmental Research and Public Health, 16*(14), 2612.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Norcross, J. C., & Lambert, M. J. (2018). Psychotherapy relationships that work III. Psychotherapy, 55(4), 303–315.

OECD. (2019). Artificial intelligence in education: Challenges and opportunities for sustainable development. OECD Publishing.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Pierce, B. S., Perrin, P. B., Tyler, C. M., McKee, G. B., & Watson, J. D. (2021). The COVID-19 telepsychology revolution: A national study of pandemic-based changes in U.S. mental health care delivery. American Psychologist, 76(1), 14–25.

Porges, S. W. (2011). The polyvagal theory: Neurophysiological foundations of emotions, attachment, communication, and self-regulation. W. W. Norton.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.

Reed, G. M., McLaughlin, C. J., & Milholland, K. (2000). Ten interdisciplinary principles for professional practice in telehealth: Implications for psychology. Professional Psychology: Research and Practice, 31(2), 170–178.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.

Riper, H., Blankers, M., Hadiwijaya, H., Cunningham, J., Clarke, S., Wiers, R., … & Cuijpers, P. (2014). Effectiveness of guided and unguided low-intensity internet interventions for adult alcohol misuse: A meta-analysis. PLoS One, 9(6), e99912.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology, 21(2), 95–103.

Rudd, M. D., Berman, A. L., Joiner, T. E., Nock, M. K., Silverman, M. M., Mandrusiak, M., & Witte, T. (2006). Warning signs for suicide: Theory, research, and clinical applications. Suicide and Life-Threatening Behavior, 36(3), 255–262.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4), 105–114.

Schueller, S. M., Muñoz, R. F., & Mohr, D. C. (2013). Realizing the potential of behavioral intervention technologies. Current Directions in Psychological Science, 22(6), 478–483.

Schüll, N. D. (2012). Addiction by design: Machine gambling in Las Vegas. Princeton University Press.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–457.

Singh, S., Drummond, C., & Ahmadi, S. (2022). AI ethics in healthcare: A systematic review. AI and Ethics, 2(1), 103–124.

Solove, D. J. (2013). Introduction: Privacy self-management and the consent dilemma. Harvard Law Review, 126(7), 1880–1903.

Stallman, H. M. (2010). Psychological distress in university students: A comparison with general population data. Australian Psychologist, 45(4), 249–257.

Steinberg, L. (2013). Adolescence (10th ed.). McGraw-Hill.

Sue, D. W., & Sue, D. (2015). Counseling the culturally diverse: Theory and practice (7th ed.). Wiley.

The New York Times. (2024, 27 de febrero). Una madre demanda a Character.AI alegando que el chatbot llevó a su hijo al suicidio. The New York Times.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

Torous, J., & Roberts, L. W. (2017). Needed innovation in digital health and smartphone applications for mental health: Transparency and trust. JAMA Psychiatry, 74(5), 437–438.

Torous, J., Lipschitz, J., Ng, M., & Firth, J. (2020). Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. *Journal of Affective Disorders, 263*, 413–419.

Twenge, J. M. (2017). iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy—and completely unprepared for adulthood. Atria Books.

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO.

U.S. Food and Drug Administration. (2019). Software as a Medical Device (SaMD): Clinical evaluation. FDA.

van der Hart, O., Nijenhuis, E. R., & Steele, K. (2006). The haunted self: Structural dissociation and the treatment of chronic traumatization. W. W. Norton.

Vallor, S. (2016). Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press.

Vayena, E., Dzenowagis, J., Brownstein, J. S., & Sheikh, A. (2018). Policy implications of big data in the health sector. Bulletin of the World Health Organization, 96(1), 66–68.

Vayena, E., Salathé, M., Madoff, L. C., & Brownstein, J. S. (2015). Ethical challenges of big data in public health. PLoS Computational Biology, 11(2), e1003904.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99.

Wampold, B. E., & Imel, Z. E. (2015). The great psychotherapy debate: The evidence for what makes psychotherapy work (2nd ed.). Routledge.

Williams, D. R., & Mohammed, S. A. (2009). Discrimination and racial disparities in health: Evidence and needed research. Journal of Behavioral Medicine, 32(1), 20–47.

Winfield, A. F. (2019). Ethical standards in robotics and AI. Nature Electronics, 2(2), 46–48.

World Economic Forum. (2020). The future of jobs report 2020. Centre for the New Economy and Society.

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. WHO.

World Medical Association. (2018). WMA statement on augmented intelligence in medical care. World Medical Association.

Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. CyberPsychology & Behavior, 1(3), 237–244.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.