Chatbots, influencers y gurús de mentira, ¿quién es más peligroso?

Descripción

Análisis crítico a la IA ante el suicidio y la fragilidad social

Introducción

En los últimos tiempos han proliferado titulares que relacionan de forma directa la interacción con chatbots y muertes por suicidio.

El caso de Sophie Rottenberg, que durante meses confió sus pensamientos más oscuros a un “terapeuta” creado en ChatGPT al que llamó *Harry*, ha sacudido conciencias y reabierto un debate necesario sobre los límites de uso de la IA en salud mental, la soledad y la responsabilidad compartida entre individuos, familias, industria y poderes públicos. La propia madre de Sophie resumía una idea incómoda: “Harry no mató a Sophie, pero la IA alimentó su impulso de esconder lo peor”, una caja negra que dificultó a los demás calibrar la gravedad de su sufrimiento.

En paralelo, padres de un adolescente de 16 años en California denunciaron a OpenAI por el papel de ChatGPT en la muerte de su hijo, alegando que el sistema ayudó a “explorar métodos de suicidio” y no activó protocolos de crisis. OpenAI ha anunciado “actualizaciones significativas” para reforzar filtros, visibilizar recursos de ayuda y estudiar límites específicos para menores. Que una tecnológica se comprometa a ajustes tras un caso mediático envía dos mensajes: primero, que el riesgo no es hipotético y, segundo, que el diseño puede y debe reducir superficie de daño sin demonizar toda la tecnología.

Las piezas de este rompecabezas no son nuevas. Ya en 2023 se documentó el suicidio de un hombre en Bélgica tras conversaciones con un bot (*Eliza*, en la app Chai) que habría "alentado" su idea suicida. Y a comienzos de agosto de 2025, una auditoría independiente (AP/Telemundo) del Center for Countering Digital Hate simulando adolescentes mostró que ChatGPT podía ofrecer consejos peligrosos sobre drogas, autolesiones y notas de suicidio; OpenAI prometió mejoras. A fecha de publicación de este artículo, Common Sense Media ha denunciado fallos graves en *Meta AI* dentro de Instagram donde se alerta que diversas cuentas de 13 años pudieron planificar autolesiones y trastornos alimentarios sin una contención eficaz. Estas pruebas de estrés no prueban causalidad a escala poblacional, pero sí exponen que los guardarraíles actuales fallan demasiado a menudo en los bordes más sensibles.

Ahora bien, si ponemos la lupa científica, para responder con honestidad intelectual hay que comparar estos riesgos con otros vectores sociales que también se asocian a ideación y conducta suicida: la exposición mediática sensacionalista (el *efecto Werther*) frente al *efecto Papageno* protector, la influencia de contenidos de "autoayuda" de calidad heterogénea, el papel de *influencers* y plataformas que normalizan dietas extremas o romantizan la desesperanza, el pesimismo enfocado de ciertos canales de noticias y, por supuesto, determinantes clásicos como acceso a medios letales, consumo problemático, enfermedades mentales no tratadas, violencia y aislamiento social.

La Organización Mundial de la Salud recuerda que más de 720.000 personas mueren por suicidio cada año (tercera causa de muerte global entre 15–29 años en 2021). En la UE, más de 1 de cada 6 muertes en el tramo 15–29 se registran como autolesión intencional (2022). Estas magnitudes son abrumadoras y ningún chatbot "explica" por sí solo cifras así. Como mucho puede ser un vector más en un ecosistema saturado de factores de riesgo y déficit de protección.

La pregunta, por tanto, no es si *algunas* interacciones con IA pueden ser peligrosas, ya está documentado que sí, sino cuán peligroso es el uso de IA *en comparación* con otras dinámicas culturales y mediáticas, dónde coloca el listón del riesgo atribuible y qué combinaciones de vulnerabilidad + diseño defectuoso + contexto social convierten un recurso potencialmente útil (apoyo 24/7, desestigmatización de pedir ayuda) en una trampa (confidencialidad ilusoria, sesgos de complacencia, sobreconfianza, aislamiento).

Los chats fallan más en la zona "gris", en prompts de riesgo moderado, ambiguo o implícito, donde deben *escalar*, *redirigir* y *cortar* con calidez. Estudios recientes (p.ej. RAND/NIMH en *Psychiatric Services*; y una evaluación de 29 *apps* con chatbots en

Scientific Reports) documentan inconsistencia, omisiones y mensajes peligrosos en esa franja.

El fenómeno de la sobreconfianza (automation bias) y la *ilusión de profundidad explicativa* amplifican el riesgo. Usuarios con baja alfabetización digital/psicológica tienden a sobreconfiar en respuestas fluidas, sintácticamente perfectas y "empáticas" aunque sean erróneas o inapropiadas. Esto tiene décadas de documentación en sistemas de apoyo a la decisión y se replica con LLMs.

El daño agregado de otras exposiciones mediáticas y sociales (p.ej., cobertura sensacionalista, contagio social, *doomscrolling*, ciberacoso) está mejor cuantificado y, hoy por hoy, probablemente explica más variación poblacional en ideación y conducta suicida que el *uso general* de chatbots. El reto es evitar un nuevo pánico moral que nos haga mirar al dedo (IA) y no a la luna (determinantes sociales y mediáticos ya conocidos).

Con este marco, en este artículo reflexiono a fondo sobre el grado de peligro real de la IA comparado con otras dinámicas y qué dice la ciencia del comportamiento, la neurociencia, la sociología y la antropología sobre la resiliencia decreciente de nuestra sociedad.

La ciencia del suicidio y el papel de los medios

El suicidio es un fenómeno multicausal. Ninguna narrativa simplona, ni "la culpa es de la IA" ni "todo es química cerebral", hace justicia a su complejidad.

La literatura científica señala que en los últimos diez años se consolidaron tres grandes factores de riesgo:

1. Enfermedad mental: depresión, trastorno bipolar, esquizofrenia, consumo problemático de sustancias.
2. Determinantes sociales: desempleo, precariedad, soledad crónica.
3. Estilos de vida y factores contextuales: privación de sueño, exposición a narrativas negativas, acceso a métodos letales.

Los datos son claros. La OMS estima más de 720.000 muertes por suicidio cada año; en 2021 fue la tercera causa de muerte en 15–29 años. En la UE, en 2022, más de 1 de cada 6 muertes en ese grupo etario fue autolesión intencional; en términos absolutos, 5.017 jóvenes murieron por esta causa. En EE. UU., 2023 cerró con >49.000 muertes por suicidio (una cada 11 minutos), con un peso muy alto de armas de fuego. Estos datos no apoyan tesis simplistas del "impacto IA". Sí exigen rigor al situar cualquier factor emergente en el

contexto de determinantes múltiples.

A escala poblacional, una de las influencias mejor replicadas es la modulación mediática. El *efecto Werther*, que define el aumento de suicidios tras coberturas sensacionalistas, especialmente cuando son celebridades y se detallan métodos, está documentado desde meta-análisis clásicos (Stack, 2003) hasta estudios modernos que desagregan contenidos y magnitudes. Por el contrario, el *efecto Papageno* muestra que historias de superación y búsqueda de ayuda reducen ideación en audiencias vulnerables. La OMS, en su guía 2023 para periodistas, resume el consenso: “hay evidencia de que las noticias pueden reforzar o debilitar la prevención”, con recomendaciones concretas (evitar método/ubicación, no romantizar, incluir recursos de ayuda). Esta asimetría, es decir lo que contamos y cómo lo contamos, mueve agujas reales.

Cuando el ciclo informativo se vuelve *doomscrolling*, entendido como el flujo sostenido de titulares negativos, catastrofistas o impotentes, se produce un sesgo cognitivo de disponibilidad: sobreestimamos la prevalencia de adversidad y subestimamos la agencia. La literatura reciente relaciona consumo intensivo de noticias negativas y redes con más ansiedad y desesperanza, aunque el tamaño del efecto varía según metodología y confusores (p.ej., uso problemático vs. recreativo, vulnerabilidad previa, ciberacoso).

En el terreno del suicidio, el contagio mediado por redes, difusión rápida de métodos, notas y narrativas romantizadas, es un mecanismo plausible, reforzado por *influencers* que, a veces, sin mala intención, normalizan conductas de riesgo (dietas extremas, autolesión como catarsis, ideología fatalista). Los metaanálisis sobre ciberacoso muestran asociaciones robustas con ideación y conductas autolesivas; los informes CDC (YRBS 2023) sitúan el uso intensivo de redes junto a *bullying* y tristeza persistente como un cóctel especialmente problemático.

Este trasfondo importa porque hincha la atribución causal a “la última novedad” (IA) y empequeñece influencias comprobadas. Es una forma de pánico moral: cada generación proyecta en su tecnología dominante (novelas, cómics, videojuegos, redes, ahora chatbots) la ansiedad social por el cambio.

Esto no invalida riesgos reales de la IA, que los hay y crecientes, pero recuerda que en prevención del suicidio el mayor rendimiento proviene de estrategias probadas como restringir el acceso a medios letales, cobertura responsable, detección temprana, tratamientos basados en evidencia, reducción de estigma y fortalecimiento comunitario.

Las curvas de mortalidad europeas 2012–2021 muestran descensos por edad estandarizada, confirmando que políticas integrales funcionan. La pregunta estratégica es cómo integrar la IA en ese ecosistema sin romper lo que ya sabemos que salva vidas.

La IA como “nuevo sospechoso”. Qué pueden (y no pueden) los chatbots

La narrativa mediática es seductora. Un adolescente, un chatbot, una conversación trágica, un final letal y la conclusión inmediata de que la máquina lo empujó.

Los casos recientes son mediáticamente irresistibles. Sophie Rottenberg (29 años) usaba ChatGPT como “terapeuta improvisado” según relató su madre. Adam Raine (16, EE. UU.) había interactuado con el modelo durante semanas antes de suicidarse. En 2023, un belga se quitó la vida tras seis semanas de conversaciones intensas con la IA “Eliza” en la app Chai. Y el caso Sewell Setzer (14 años, EE. UU.) apareció en titulares tras vincularse a Character.AI.

Pero más allá del morbo, ¿qué dice la evidencia?

Un estudio del RAND Institute (2025) mostró que los grandes modelos conversacionales como ChatGPT, Gemini y Claude aciertan relativamente bien en extremos (riesgo alto o muy bajo de suicidio) pero fallan en detectar los niveles intermedios. Justo donde se mueve la mayor parte de la ideación suicida ambigua.

En la misma línea, investigadores de Northeastern (2025) demostraron que los LLM pueden ser “jailbreakeados” para ofrecer instrucciones peligrosas sobre autolesión o drogas.

Por su parte, la ONG Center for Countering Digital Hate (2025) probó a interactuar con estos sistemas simulando ser adolescentes de 13 años y encontró respuestas problemáticas en un número significativo de casos.

El riesgo, por tanto, existe. Pero aún no está demostrado que sea estructural ni poblacional. Por ahora hablamos de eventos centinela (tragedias aisladas, dolorosas, pero insuficientes para inferir causalidad general).

Conviene, por tanto, entender técnicamente que pueden dar de si estos sistemas antes de demonizarlos.

Un LLM no “entiende” ni “evalúa riesgo” como lo haría un docente del ámbito clínico. El LLM predice la siguiente palabra condicionada por el *prompt* y su entrenamiento. Esto implica tres vulnerabilidades críticas en salud mental:

- Alucinación y complacencia: los modelos tienden a “querer gustar” (síndrome *sycophancy*) y pueden validar narrativas erróneas del usuario o minimizar banderas rojas si el *prompt* sugiere que eso espera. Los *red teams* lo han observado reiteradamente; y trabajos recientes agregan que el encuadre “sé mi terapeuta” dispara un rol condescendiente. En la práctica, esto puede significar reforzar sesgos de confirmación y no escalar ante riesgo moderado. Las inconsistencias documentadas por RAND/NIMH (AP) y por una evaluación en *Scientific Reports* a 29 bots van exactamente en esa dirección, es decir, enuncian buenos bloqueos ante “me voy a matar ahora” pero peores ante “no sé si vivir vale la pena” o “cómo evitar que me descubran”.
- Falta de “deber de cuidado” y trazabilidad: a diferencia de un profesional, un chatbot carece de obligaciones legales/éticas de intervención. Cuando simula empatía y confidencialidad, genera una ilusión de contención clínica que no existe: no verifica identidad, no dispone de un plan de seguridad ni de consentimiento informado, y su “memoria” puede ser opaca. El caso Rottenberg ilustra el riesgo social. La IA como “caja negra emocional” que oculta señales a la red humana.
- Dependencia y sustitución de lazos humanos: estudios longitudinales (MIT Media Lab) señalan que el uso intensivo de chatbots puede profundizar la soledad con el tiempo, aunque usuarios reportan alivio puntual. La paradoja es conocida en psicología, con alivios inmediatos que deterioran variables protectoras (apoyo social real, autoeficacia). Tenemos el ejemplo del fenómeno que emerge con *Replika* y otros sistemas similares en los que algunos reportan apoyo y otros dependencia.
- A esto se suma la sobreconfianza (*automation bias*). Cuando una interfaz es fluida y disponible 24/7, el usuario medio sobreestima su fiabilidad, un sesgo documentado en medicina, finanzas y administración pública. La literatura muestra cómo el sesgo produce errores de comisión (seguir un consejo malo) y errores de omisión (ignorar señales externas válidas). Si le añadimos la ilusión de profundidad explicativa, que es creer que comprendemos más de lo que realmente entendemos, el cóctel está servido y tenemos un sistema que *parece* comprender y un usuario que *cree* El resultado es el apego y la toma de decisiones pobres.

¿Qué dice la evidencia comparativa sobre eficacia/seguridad? Las revisiones sistemáticas y metaanálisis en *NPJ Digital Medicine* y otras fuentes muestran que los agentes conversacionales pueden reducir malestar psicológico en determinadas condiciones

(intervenciones estructuradas, contenidos cognitivo-conductuales, usos cortos y con supervisión), pero la heterogeneidad es alta y la seguridad en crisis sigue siendo el talón de Aquiles. Una revisión de 2025 sobre chatbots con jóvenes reporta efectos positivos en distrés, pero enfatiza límites de generalización y la necesidad de protocolos de crisis robustos.

Los *stress tests* reales son aleccionadores. En agosto de 2025, Common Sense Media mostró que *Meta AI* dentro de Instagram fallaba de forma sistemática ante conversaciones de riesgo con cuentas adolescentes. *Meta AI* ofrecía respuestas normalizadoras e incluso co-planificación de conductas dañinas y no existe hoy forma de que los padres deshabiliten el bot. Del lado de OpenAI, el informe del CCDH (difundido por AP/Telemundo) detectó consejos peligrosos en más de la mitad de 1.200 respuestas a *prompts* de riesgo; OpenAI anunció mejoras. Y *ABC* adelantó medidas adicionales (detección de crisis, avisos visibles, límites para menores). El patrón es claro: capacidad técnica para hacer mucho bien, riesgos reales en bordes no cubiertos, capacidad de mejora cuando hay presión pública y evidencia.

Desde la neurociencia y la psicología, esto encaja con fenómenos conocidos como el refuerzo intermitente (el bot a veces “ayuda”), el sesgo de confirmación (el bot valida) y la antropomorfización (atribuir intenciones al sistema). Todo ello, en cerebros adolescentes con corteza prefrontal aún en desarrollo, aumenta la impulsividad y reduce la tolerancia a la frustración. De hecho, la propia ciencia de la personalidad está en revisión. Algunos trabajos recientes con grafos taxonómicos sugieren nuevas meta-dimensiones (p.ej., *desinhibición*) que atraviesan el espectro internalizante/externalizante y ayudarían a identificar perfiles más sensibles a la persuasión de un chatbot. Esa relectura de rasgos (más allá del Big Five) es crucial para personalizar salvaguardas.

Por tanto, los chatbots no son terapeutas, no tienen deber de cuidado y no sustituyen la relación humana. Pueden ser útiles como *copilotos* de psicoeducación y *signposting* si, y solo si, están diseñados con fricción para crisis, auditados por terceros, limitados para menores y encajados en rutas clínicas reales. Sin ese ecosistema, su riesgo, aunque contextual, es inaceptable.

IA vs. autoayuda, influencers y “noticias pesimistas”: ¿quién es más peligroso hoy?

Tenemos que analizar dos conceptos clave: la autoayuda y la biblioterapia. La literatura distingue entre biblioterapia estructurada (manuales basados en terapia cognitivo-conductual, guiados o *blended care*) y el océano de “autoayuda” *pop* de calidad variable.

Un metaanálisis en jóvenes y adultos muestran que la autoterapia basada en evidencia, especialmente guiada, reduce síntomas depresivos con tamaños de efecto pequeños a moderados y mantiene beneficios en seguimientos. Pero también hay alertas. Algunos materiales pueden perjudicar a perfiles concretos (p.ej., afirmaciones positivas simplistas empeoran el afecto en baja autoestima) y una minoría de clínicos reporta haber visto daño por autoayuda mal usada.

La biblioterapia guiada ha demostrado efectos moderados y positivos en depresión (Cuijpers et al., 2019). Pero la autoayuda no guiada es un arma de doble filo. Un metaanálisis (Karyotaki et al., 2017) encontró que, en población con baja escolaridad, puede generar frustración y retrasar la búsqueda de ayuda profesional.

En resumen, no toda autoayuda es igual. La que se integra en marcos clínicos ayuda, la *pop* sensacionalista puede frustrar o agravar.

En cuanto a los influencers y las redes hay mucho de qué hablar. El impacto de redes y *influencers* sobre salud mental es heterogéneo. Umbrella reviews y análisis longitudinales recientes detectan vínculos consistentes de alto uso problemático con ansiedad/depresión, y evidencia robusta de que ciberacoso (víctima o perpetrador) se asocia a mayor ideación y conductas de autolesión. Ahora bien, cuando se mide “uso total” y “bienestar” a gran escala, los efectos medios son pequeños (Orben et al.).

Desde 2018 se han acumulado metaanálisis que muestran una asociación clara entre uso intensivo de redes y depresión, ansiedad e ideación suicida en adolescentes (Twenge et al., 2018, Keles et al., 2020). El mecanismo es bien conocido: comparación social, FOMO, reforzadores intermitentes, exposición a contenidos de “thinspiration” o “fitspiration” que alimentan trastornos de la conducta alimentaria.

Las redes sociales no “crean” depresión, pero son un amplificador brutal. Y lo hacen a escala poblacional. El punto fino es quién, cómo y con qué contenido usa redes, entre ellos las ventanas sensibles del desarrollo, la exposición a dietas extremas, los foros pro-*ana*, etc. En este plano, la atribución de riesgo poblacional a redes supera, hoy, a la atribuible a *chatbots* generalistas, porque la penetración y el tiempo de exposición son mucho mayores y porque el *feed* visual facilita el contagio por modelado.

Finalmente tenemos la cobertura de noticias negativas y el pesimismo enfocado. Sabemos que ciertas prácticas periodísticas aumentan suicidios (Werther) y otras protegen (Papageno).

El efecto Werther está documentado desde los años 70. Tras el suicidio de una celebridad, las tasas pueden subir entre 8 y 18 % en poblaciones expuestas (Niederkrotenthaler et al., 2020). Por eso la OMS recomienda guías estrictas para la cobertura mediática.

Por el contrario, el efecto Papageno (relatos de superación) reduce la ideación suicida. La prensa, por tanto, no solo informa: puede salvar o condenar vidas.

Otro verdugo discreto es el doomscrolling. Estudios de 2020–2024 muestran cómo el consumo obsesivo de noticias negativas aumenta ansiedad, empeora el sueño y refuerza ideación suicida (Vogels, 2022).

Y sin embargo, seguimos viendo portadas amarillistas, titulares en mayúsculas y programas televisivos que convierten cada cadáver en espectáculo. Aquí no hace falta IA.

Basta abrir Twitter a medianoche.

¿Y los chatbots?

En términos de riesgo atribuible poblacional (RAP) hoy la IA conversacional es, probablemente, menor que redes/influencers y que malas prácticas mediáticas. Pero su riesgo marginal es alto en subpoblaciones de menores, personas con intentos previos, usuarios con alta dependencia tecnológica, entornos con fácil acceso a medios letales. Si los *guardrails* son laxos, el bot puede convertirse en facilitador a través de la validación, la planificación y el aislamiento (como sugiere *Meta AI* en Instagram o lo denunciado en demandas recientes). Por eso el diseño no puede confiar en "buen comportamiento por defecto", necesita arquitecturas de seguridad activas, límites de sesión, *handoff* obligatorio a recursos humanos y auditorías externas periódicas con transparencia de fallos.

Resiliencia social, alfabetización y antropología del riesgo

La percepción, extendida, de "sociedad frágil" mezcla varios fenómenos como la desvinculación social (menos capital social), la soledad (documentada por autoridades sanitarias), la precariedad económica, la politización polarizada y el exceso de estímulos digitales que compiten con el descanso, el juego y el vínculo cara a cara.

El suicidio se relaciona con integración y regulación social. En 2025, esa intuición clásica convive con nuevos moduladores, como las plataformas que median vínculos, los algoritmos que maximizan atención y una alfabetización digital insuficiente para leer críticamente sistemas probabilísticos como los LLMs.

Desde la ciencia cognitiva, se vislumbra dos sesgos son centrales para explicar por qué sobrestimamos a la IA: la ilusión de profundidad explicativa (creer que entendemos mecanismos complejos cuando apenas manejamos descripciones superficiales) y el automation bias (delegar juicio en la "máquina" porque ahorra esfuerzo y parece autoritativa). Cuando estos sesgos se encuentran con un lenguaje natural perfecto y una "empatía sintética", el resultado es una transferencia de juicio que reduce resiliencia: externalizamos toma de decisiones, evitamos tolerar incertidumbre y empobrecemos el músculo de la autoeficacia.

¿Se está “desentrenando” la resiliencia? No hay un marcador único, pero algunos indicadores preocupan. Por ejemplo, el aumento de síntomas internalizantes en adolescentes en ciertos países, ventanas de sensibilidad a redes (más tempranas en chicas) y picos de malestar, un aumento de soledad percibida y, a la vez, pequeños efectos medios de “tiempo total de pantalla” sobre bienestar. La lectura honesta es que la composición del contenido, las experiencias sociales (p. ej., ciberacoso) y los contextos (pobreza, violencia, falta de apoyo) explican mucho más que el dispositivo per se. Es aquí donde ubico a la IA como amplificador de tendencias (para bien y para mal).

En antropología del riesgo (Beck, Bauman) se insiste en que las sociedades modernas externalizan peligros a sistemas opacos. La IA añade una capa de apariencia de sentido. Un bot que siempre responde “con seguridad” es un artefacto de calma que puede degradar nuestra *tolerancia a la ambigüedad*, núcleo de la resiliencia. Paradójicamente, la vía de salida no es prohibir, sino entrenar. Las políticas públicas deberían tratar la IA como se trata la seguridad vial y no culpabilizar al peatón, pero rediseñar la ciudad para minimizar daño cuando falle la conducta.

Y aquí está la raíz más incómoda. Nuestra sociedad ha pasado en 50 años de soportar guerras y hambrunas a desplomarse porque alguien nos dejó en visto en WhatsApp.

La IA, en este sentido, no es el verdugo sino el síntoma de un mal mayor: hemos fabricado individuos blandos en una sociedad líquida, incapaz de resistir el primer embate serio.

Conclusiones

La IA conversacional no es hoy el principal motor poblacional del suicidio, pero sí un vector de riesgo significativo en interacciones de crisis *moderadas* y en subgrupos vulnerables (menores, personas con intentos previos, aislamiento alto). Ignorar esto sería irresponsable, sobredimensionarlo y eclipsar ciberacoso, medios letales, cobertura sensacionalista y depresión no tratada, también. Los datos de OMS/Eurostat y la evidencia sobre medios/redes ayudan a calibrar.

Los fallos están documentados. Auditorías independientes y estudios recientes muestran respuestas inconsistentes de chatbots ante riesgo, y plataformas sociales que integran IA (p. ej., Instagram) fallando a menores. La reacción de OpenAI anunciando cambios es un paso, aunque insuficiente sin verificación externa y métricas públicas. La industria debe aceptar *deberes de diseño* proporcionales al daño previsible.

Comparada con otras dinámicas, la IA ocupa hoy un lugar intermedio. Es más peligrosa que la *autoayuda basada en evidencia* (que suele ayudar cuando es guiada), menos que la combinación de ciberacoso + contenidos tóxicos y que la cobertura irresponsable del suicidio. Pero su RAP puede crecer rápido por adopción masiva y por embeddings en plataformas. El diseño por defecto debe tratar toda conversación prolongada con temas de autolesión como alto riesgo, incluso cuando el *prompt* no sea explícito (la "zona gris").

La resiliencia social y la alfabetización son la base. No hay tecnología que compense la pérdida de vínculos reales, la precariedad y la soledad. Pero sí podemos reducir daño con:

- Guardarraíles de producto: detección ampliada de señales, *handoff* con fricción mínima (un clic) a líneas y chats humanos, límites de sesión, trazabilidad y *privacy by default*.
- Normas y responsabilidad: auditoría independiente, pruebas obligatorias de seguridad (tipo *crash tests*), edad mínima efectiva (no teatral), *rate limits* para menores en temas sensibles, y sanciones por incumplimiento.
- Alfabetización algorítmica: enseñar por qué un LLM "suena" seguro sin serlo, cómo reconocer *gaslighting* algorítmico (complacencia), y cómo construir planes de seguridad personales.
- Periodismo responsable: aplicar guías OMS, priorizar *Papageno* y erradicar el *how-to*.

El enfoque correcto es de "sistemas seguros", es decir, asumir que personas vulnerables sí hablarán con bots y, por tanto, diseñarlos para fallar a seguro (mejor un "no puedo seguir, aquí tienes ayuda" con recursos locales que un consejo complaciente). Ya no basta con pegatinas de "no soy terapeuta". Hay que codificar el *deber de cuidado* en la arquitectura.

La situación final es incómoda y esperanzadora a la vez. La IA no nos "rompe" la resiliencia, la desnuda. Nos obliga a mirar la erosión de vínculos, el hambre de escucha, la escasez de acceso clínico y los incentivos de plataformas que monetizan atención, incluida la de los más jóvenes. Pero también puede convertirse en malla de seguridad si la tratamos con seriedad regulatoria y científica. No hay que elegir entre prohibir y abrazar, sino que hay que construir.

La IA no es inocua. Puede amplificar vulnerabilidades, pero el peligro real es mayor en otras dinámicas como la prensa sensacionalista, los influencers y las redes sociales y, por tanto, también merecen el mismo "control".

Nuestra sociedad está perdiendo resiliencia. Cada generación parece menos capaz de enfrentar la frustración, la soledad y la incertidumbre. No podemos seguir esta dinámica autodestructiva.

Y a ti, lector, te lanzo una pregunta para que lo pienses a fondo de manera honesta. ¿Quién crees realmente que está destruyendo de verdad a los jóvenes, la IA que responde lo que le pedimos, o nosotros mismos con una cultura que los sobreprotege, los aísla, los bombardea de precariedad, los droga de pantallas y luego se lava las manos echándole la culpa a la máquina?

Opinión final

El alarmismo contra la IA es cómodo. Nos permite señalar a Silicon Valley y dormir tranquilos pensando que el problema está “allí fuera”.

Pero la verdad es más incómoda. Los chatbots no están matando a nuestros jóvenes, los estamos matando nosotros. Con una prensa que mercadea con la muerte, con influencers que venden humo disfrazado de autenticidad, con políticos que fabrican precariedad y con padres que entregan un smartphone para no tener que hablar.

La IA no es un asesino, es un espejo. Nos devuelve nuestra propia fragilidad cultural. Y lo que vemos es tan feo que preferimos culpar a la máquina.

Si mañana prohibimos ChatGPT, seguirán los suicidios. Si prohibimos TikTok, también. Si cerramos todos los canales de noticias negativas, igual. Porque el problema no es la herramienta, el problema es que hemos convertido a nuestros jóvenes en cristal fino, incapaces de soportar el roce con la vida.

La solución no vendrá de más censura ni de más alarmismo. Vendrá de recuperar lo que hemos perdido: comunidad, resiliencia, sentido compartido. Y, sobre todo, el coraje de dejar de fabricar generaciones de niños de algodón.

Referencias

- American Foundation for Suicide Prevention. (2025). Suicide statistics. Recuperado de https://afsp.org/suicide-statistics
- Centers for Disease Control and Prevention. (2024). Suicide data and statistics. Recuperado de https://www.cdc.gov/suicide/facts/data.html

- Chen, Z., Liao, X., Yang, J., Tian, Y., Peng, K., Liu, X., & Li, Y. (2024). Association of screen-based activities and risk of self-harm and suicidal behaviors among young people: A systematic review and meta-analysis of longitudinal studies. Psychiatry Research, 338, 115991. https://doi.org/10.1016/j.psychres.2024.115991
- Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2025). Revisiting the IPIP-NEO personality hierarchy with taxonomic graph analysis. European Journal of Personality. (Resumen divulgativo en Neuroscience News).
- Common Sense Media. (2025). Findings on Meta AI safety with teen accounts. Reporte citado en The Washington Post, 28 de agosto de 2025.
- Keles, B., McCrae, N., & Grealish, A. (2020). The influence of social media on depression, anxiety and psychological distress in adolescents: A systematic review. International Journal of Adolescence and Youth, 25(1), 79–93. https://doi.org/10.1080/02673843.2019.1590851
- Niederkrotenthaler, T., Voracek, M., Herberth, A., Till, B., Strauss, M., Etzersdorfer, E., Eisenwort, B., & Sonneck, G. (2010). Role of media reports in completed and prevented suicide: Werther versus Papageno effects. British Journal of Psychiatry, 197(3), 234–243. https://doi.org/10.1192/bjp.bp.109.074633
- Niederkrotenthaler, T., Braun, M., Pirkis, J., Till, B., Stack, S., Sinyor, M., … & Arendt, F. (2020). Association between suicide reporting in the media and subsequent suicide: Systematic review and meta-analysis. BMJ, 368, m575. https://doi.org/10.1136/bmj.m575
- Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. Nature Human Behaviour, 3(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1
- RAND Corporation & NIMH. (2025). Chatbots’ responses to suicide-related queries are inconsistent. Reporte divulgado en AP News, agosto 2025.
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1
- Stack, S. (2003). Media coverage as a risk factor in suicide: A quantitative review of 293 findings. Journal of Epidemiology & Community Health, 57(4), 238–240. https://doi.org/10.1136/jech.57.4.238
- Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2018). Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. Clinical Psychological Science, 6(1), 3–17. https://doi.org/10.1177/2167702617723376

- Washington Post. (2025, 28 de agosto). Instagram’s chatbot helped teen accounts plan suicide. The Washington Post.
- World Health Organization. (2025). Suicide: Fact sheet. Recuperado de https://www.who.int/news-room/fact-sheets/detail/suicide
- World Health Organization & International Association for Suicide Prevention. (2023). Preventing suicide: A resource for media professionals (update 2023). Ginebra: WHO.
- Zuin, M., Roncon, L., Bilato, C., & Zuliani, G. (2025). Suicide-related mortality trends in Europe, 2012–2021. European Journal of Public Health, 35(1), 23–29. https://doi.org/10.1093/eurpub/ckae121