
PUBLICACIONES DE LA
ACADEMIA NACIONAL DE
MEDICINA DE MÉXICO

# ACTUALIDADES EN INTELIGENCIA ARTIFICIAL

Dr. Raúl Carrillo Esper
Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra

Número 1

# Inteligencia artificial agente: de la automatización inteligente a la acción autónoma orientada a objetivos

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra

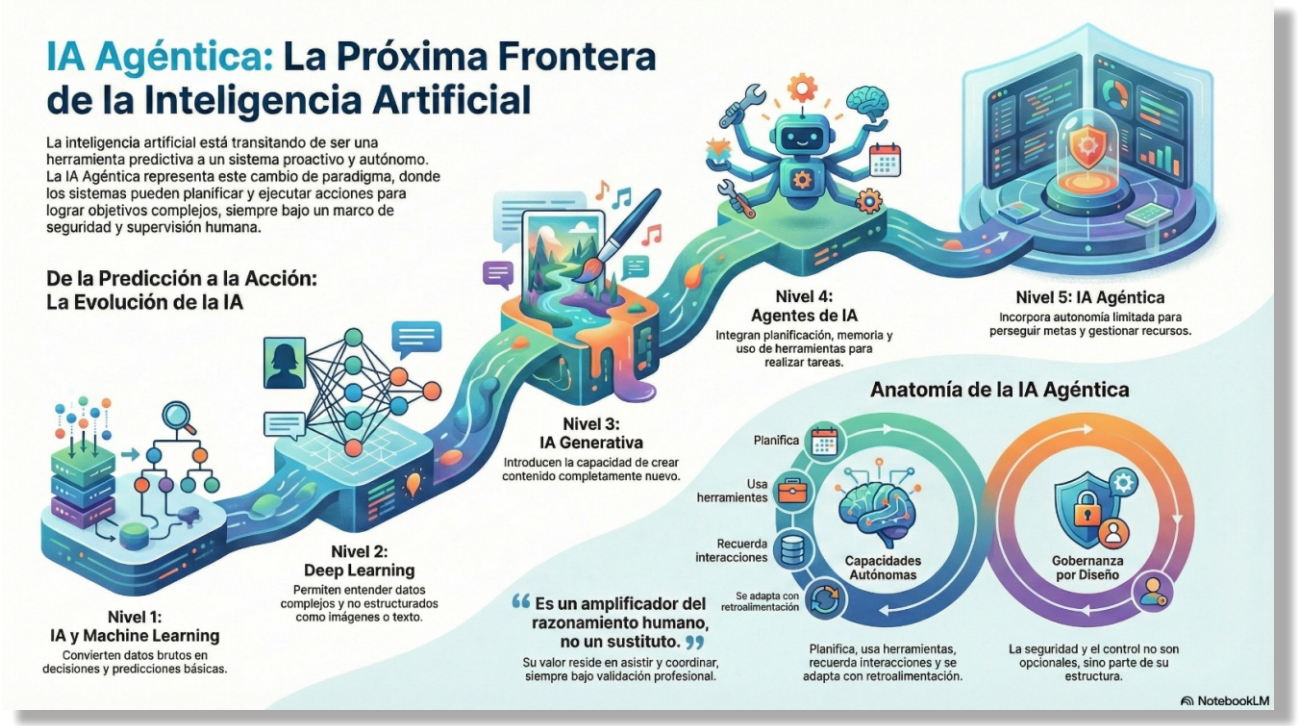
Médicos Internistas
Universidad de Guadalajara
Instituto Mexicano del Seguro Social (IMSS)
Colegiados y Certificados (CMMI)
Fundadores del TICC
28 de enero de 2026

**Introducción**

La inteligencia artificial (IA) ha experimentado una evolución acelerada durante la última década, transitando desde modelos predictivos clásicos hasta sistemas generativos capaces de producir contenido complejo y multimodal. Sin embargo, en años recientes ha emergido un nuevo paradigma: la Inteligencia Artificial Agéntica (Agentic AI), caracterizada por la capacidad de planificar, ejecutar acciones, perseguir objetivos y adaptarse dinámicamente a entornos cambiantes, todo ello bajo distintos niveles de supervisión humana.

El marco conceptual presentado en la imagen *Agentic IA: A Complete Framework* sintetiza esta evolución mediante una arquitectura en capas que va desde AI & Machine Learning, pasando por Deep Learning, Generative AI, AI Agents, hasta culminar en Agentic AI. Este modelo no solo describe avances técnicos, sino que introduce un cambio profundo en la relación entre humanos y sistemas inteligentes: de herramientas pasivas a entidades computacionales con agencia limitada y gobernanza explícita.

Dada la creciente aplicación de estas tecnologías en dominios de alto impacto como la medicina, la educación médica y la gestión de sistemas complejos resulta imprescindible analizar este marco desde una perspectiva crítica, ética y operativa, priorizando la seguridad, la trazabilidad y el juicio humano. El objetivo de este artículo es analizar de forma estructurada el marco de Agentic AI, evaluando su evolución técnica, capacidades emergentes y sus implicaciones prácticas, especialmente en contextos clínicos y educativos.



**Métodos**

Se realizó un análisis conceptual y estructural cualitativo del marco visual *Agentic IA: A Complete Framework*, utilizando una metodología de análisis por capas y dominios funcionales. El procedimiento incluyó:

1. Descomposición del marco en cinco niveles evolutivos:
   - IA y aprendizaje automático
   - Aprendizaje profundo
   - IA generativa
   - Agentes de IA
   - Agente de IA
2. Identificación de capacidades técnicas clave en cada nivel (razonamiento, generación, planificación, autonomía, gobernanza).
3. Análisis funcional de los componentes transversales:
   - Capacidades del agente
   - Gestión del agente
   - Interfaces y salidas
   - Gobernanza y seguridad
4. Interpretación aplicada del marco en escenarios de medicina y educación médica, utilizando principios de:
   - Humano en el circuito
   - Seguridad clínica
   - Toma de decisiones basada en evidencia

El análisis se desarrolló bajo un enfoque no experimental, de tipo analítico-narrativo, con énfasis en coherencia conceptual, aplicabilidad práctica y alineación con principios éticos y regulatorios ampliamente aceptados.

**Resultados**

1. Evolución funcional de la IA.

El marco evidencia una progresión clara desde sistemas orientados a datos hasta sistemas orientados a objetivos:

- AI & Machine Learning: convierten datos en decisiones mediante modelos predictivos y clasificadores.
- Deep Learning: permiten abstracción profunda y manejo de datos no estructurados.
- Generative AI: introduce la capacidad de crear contenido nuevo (texto, imagen, audio, video).
- AI Agents: integran planificación, uso de herramientas, memoria y autoevaluación.
- Agentic AI: incorpora autonomía limitada, encadenamiento de metas y gestión de recursos.

2. Capacidades emergentes de los sistemas agénticos.

Se identificaron como capacidades centrales de la Agentic AI:

- Planificación jerárquica (task decomposition, goal chaining)
- Persistencia de estado y memoria
- Retroalimentación continua y mecanismos de rollback
- Uso autónomo de herramientas
- Observabilidad y trazabilidad de acciones

Estas capacidades permiten a los sistemas no solo generar recomendaciones, sino ejecutar flujos de trabajo completos, bajo reglas predefinidas.

3. Gobernanza y control como componentes estructurales.

El marco incorpora explícitamente mecanismos de:

- Supervisión humana (human-in-the-loop)
- Gestión de errores y recuperación
- Límites de acción (guardrails)
- Cumplimiento normativo y seguridad

La gobernanza no aparece como un elemento accesorio, sino como un componente estructural indispensable para la operación segura de sistemas autónomos.

4. Aplicaciones en salud y educación médica.

En contextos clínicos y educativos, el marco sugiere que la Agentic AI puede:

- Apoyar la coordinación de procesos clínicos complejos
- Reducir carga cognitiva del profesional
- Estandarizar flujos de atención
- Facilitar tutoría adaptativa y evaluación formativa

Sin sustituir en ningún caso el juicio clínico humano.

**Discusión**

El análisis del marco Agentic AI: A Complete Framework revela que la Agentic AI representa un cambio de paradigma, más que una simple evolución tecnológica.

La transición de sistemas que responden a sistemas que actúan introduce riesgos nuevos, particularmente en entornos donde los errores tienen consecuencias clínicas, éticas o legales significativas.

Uno de los hallazgos más relevantes es que la autonomía, tal como se presenta en el marco, no es absoluta, sino condicionada por mecanismos de supervisión, gobernanza y reversibilidad. Este enfoque resulta coherente con los principios de seguridad en medicina, donde ninguna decisión crítica puede ser completamente delegada a un sistema automatizado.

Desde una perspectiva clínica, la Agentic AI debe entenderse como un amplificador del razonamiento humano, no como un sustituto. Su mayor valor reside en la organización, priorización y ejecución asistida de tareas complejas, siempre bajo validación profesional. En educación médica, su potencial radica en la tutoría cognitiva, la simulación y el acompañamiento formativo continuo.

No obstante, la implementación de Agentic AI exige marcos regulatorios claros, alfabetización digital avanzada y una cultura institucional que priorice la responsabilidad humana. Sin estos elementos, la autonomía tecnológica puede convertirse en una fuente de riesgo más que de valor.

**Conclusiones**

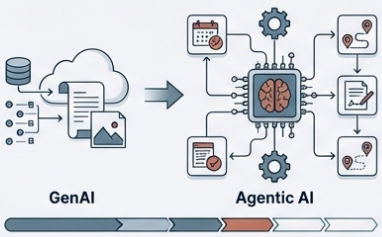
La Agentic AI constituye una evolución significativa de la inteligencia artificial, caracterizada por la capacidad de perseguir objetivos y ejecutar acciones de forma autónoma y regulada. El marco analizado proporciona una arquitectura integral
que combina capacidades técnicas avanzadas con mecanismos explícitos de gobernanza y seguridad.

En medicina y educación, su implementación debe ser prudente, supervisada y centrada en el humano, garantizando trazabilidad, reversibilidad y responsabilidad profesional. El futuro de la IA no

# Resumen Ejecutivo

## La Evolución

La IA está transicionando de generar contenido (GenAI) a ejecutar objetivos complejos (Agentic AI).

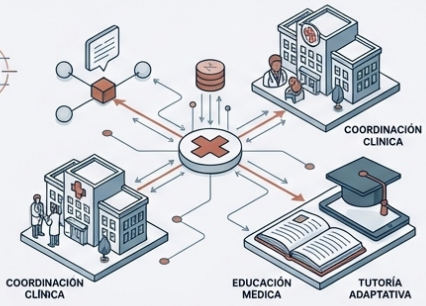



## El Marco Estructural

Se basa en una arquitectura de capas que integra planificación jerárquica, memoria persistente y uso de herramientas.

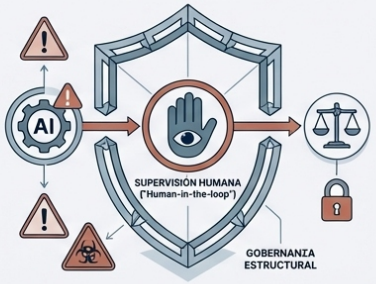

## Impacto Sectorial

Alto potencial para la coordinación clínica y la tutoría adaptativa en educación médica.



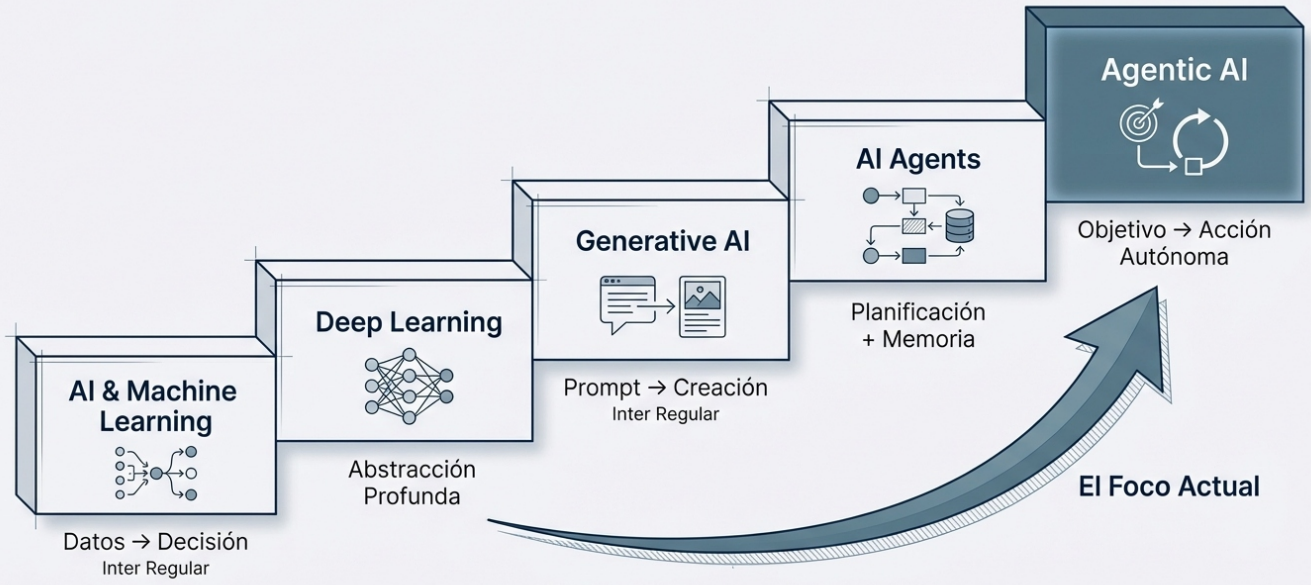

## Condición Crítica

La autonomía no es absoluta. Requiere una gobernanza estructural y supervisión humana ('Human-in-the-loop') para mitigar riesgos clínicos y éticos.




NotebookLM

# La Evolución Funcional de la Inteligencia Artificial



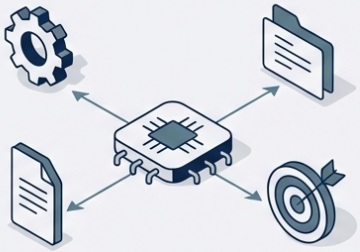

NotebookLM

# Definición y Concepto: El Salto a la Agencia

La Inteligencia Artificial Agéntica se caracteriza por la capacidad de planificar, ejecutar acciones, perseguir objetivos y adaptarse dinámicamente a entornos cambiantes.

**Herramienta Pasiva**

Responde a preguntas, espera instrucciones.



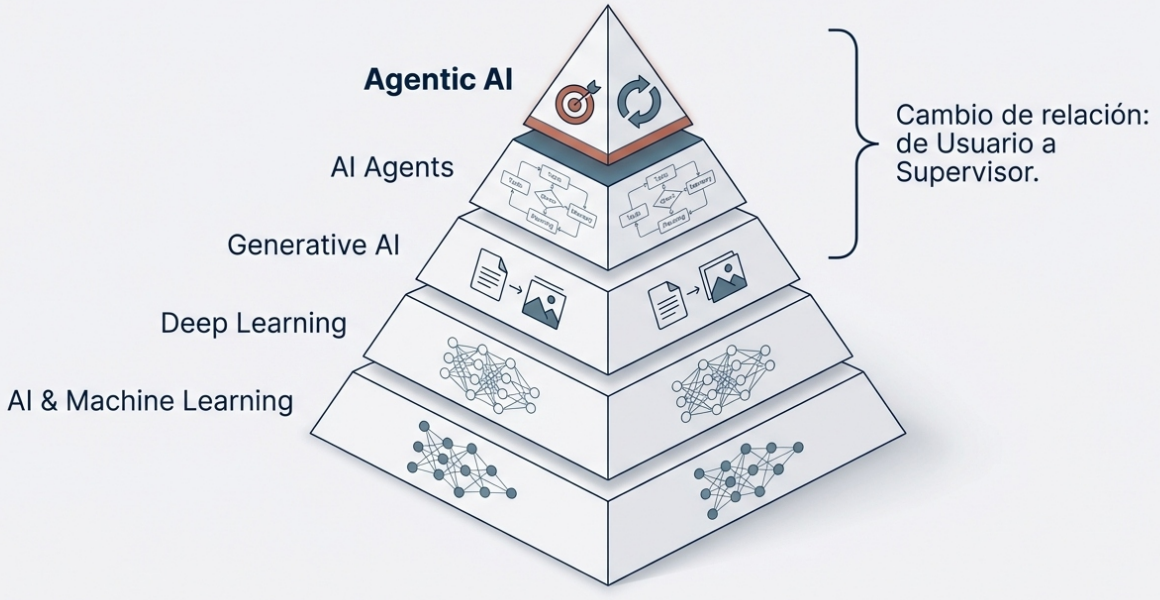
**Entidad Agéntica**

Persigue objetivos, utiliza herramientas, mantiene estado.

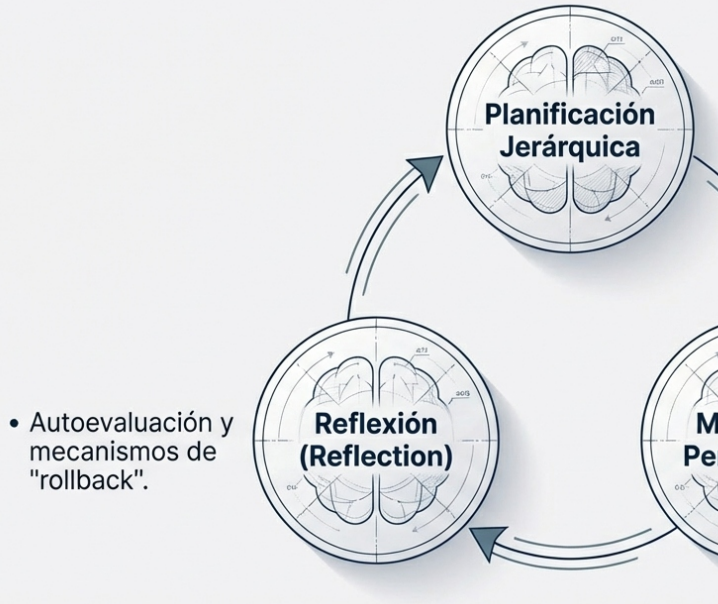
Transición de Usuario a Supervisor



# El Marco Estructural Completo



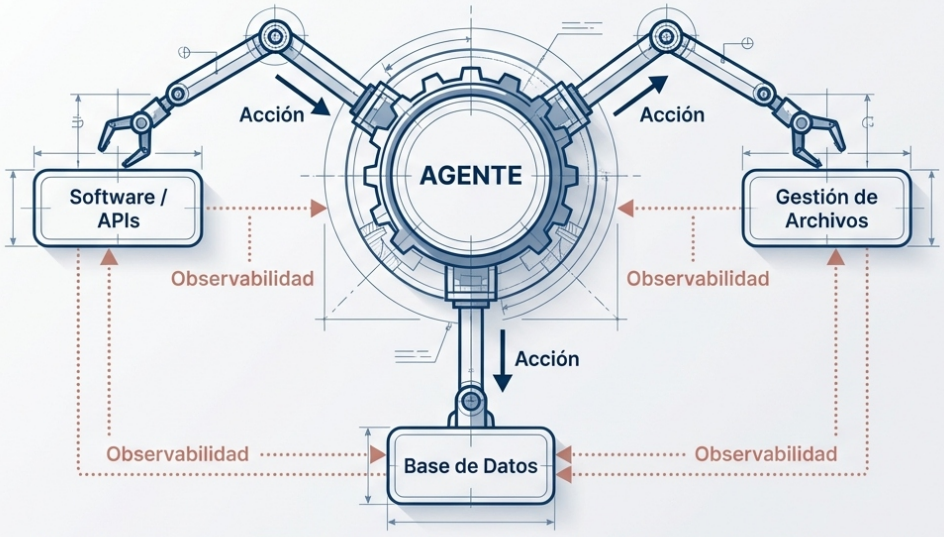
# Capacidades Centrales: El ‘Cerebro’ del Agente




- Desglose de tareas (Task Decomposition) y encadenamiento de metas (Goal Chaining).
- Mantenimiento del contexto a largo plazo.
- Autoevaluación y mecanismos de "rollback".

NotebookLM

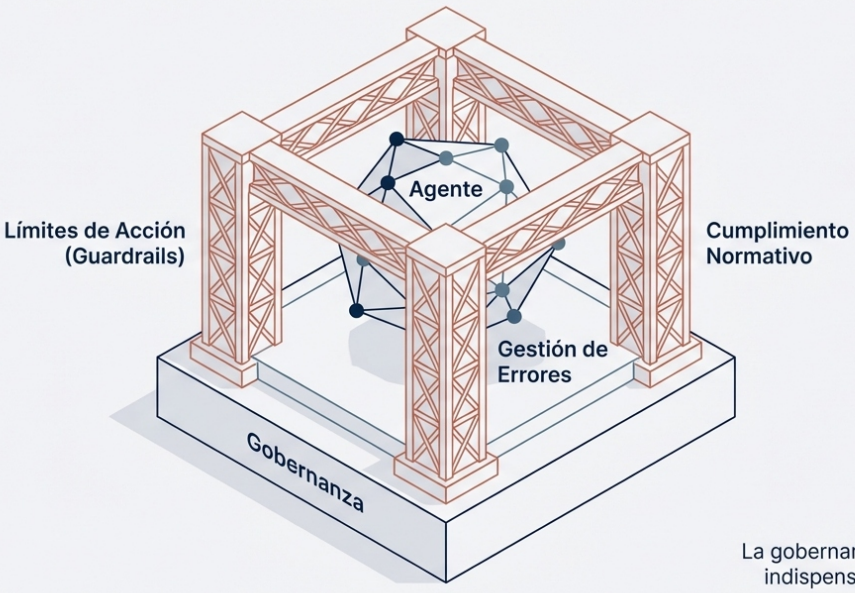
# Autonomía y Ejecución: Las ‘Manos’ del Agente




El sistema evoluciona de “Orientado a Datos” a “Orientado a Objetivos”.

NotebookLM

# La Gobernanza como Componente Estructural




La gobernanza no es accesoria; es indispensable para la operación.

NotebookLM

# El Humano en el Circuito (Human-in-the-Loop)


En medicina, ninguna decisión crítica se delega completamente.

NotebookLM

## Aplicaciones Sectoriales: Transformación Clínica

### Coordinación de Procesos

Gestión de flujos clínicos complejos y logística hospitalaria.

### Reducción de Carga Cognitiva

Automatización administrativa para permitir enfoque en el paciente.

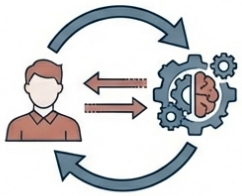
### Estandarización

Aseguramiento de adherencia a protocolos clínicos.

Sin sustituir en ningún caso el juicio clínico humano.

NotebookLM

## Aplicaciones Sectoriales: Revolución en Educación Médica



### Tutoría Adaptativa

Agentes que actúan como tutores personalizados, ajustando contenido al nivel del estudiante.

### Simulación Clínica

Generación de pacientes virtuales dinámicos que reaccionan a intervenciones.

### Evaluación Formativa

Feedback continuo y detallado sobre el razonamiento clínico.

El futuro de la formación médica, amplificado por la inteligencia artificial.

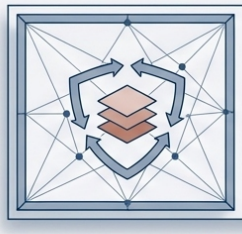
NotebookLM

# Gestión de Riesgos y Seguridad (TRiSM & AGENTSAFE)

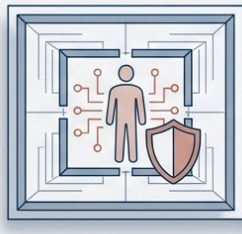
## LOS RIESGOS
Inter Medium, Deep Oxford Blue

- Alucinaciones que se convierten en acciones.
- Falta de reversibilidad.

## LOS MARCOS DE REFERENCIA
Inter Medium, Deep Oxford Blue



**TRiSM**
Trust, Risk, and Security Management.

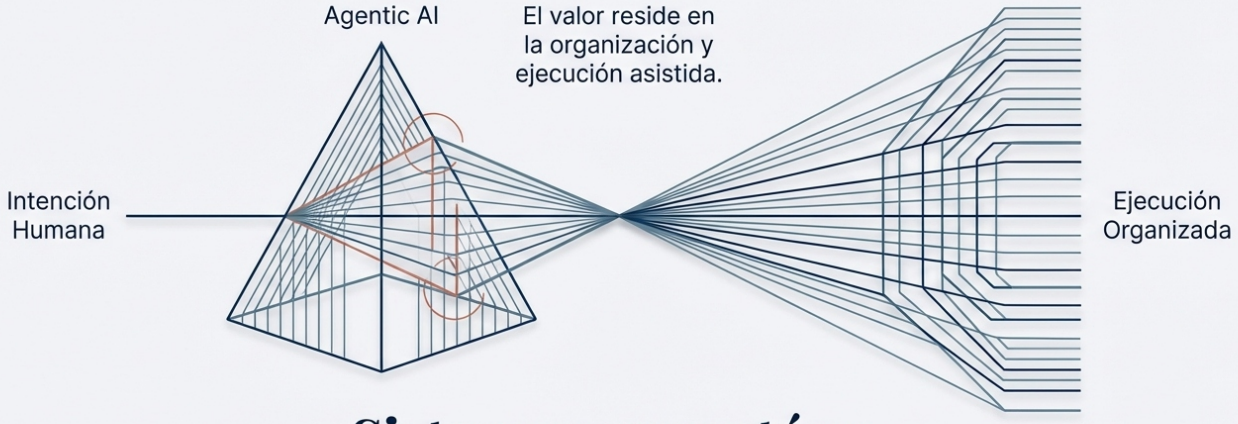


**AGENTSAFE**
Aseguramiento ético y responsabilidad operacional.

**Requisito: Botón de Deshacer (Reversibilidad).**

NotebookLM

# Discusión: El Amplificador del Razonamiento




**Sistemas que actúan,**
**“Sistemas que actúan, validados por expertos.”**

NotebookLM

# Hacia una Implementación Responsable


### 1. Priorizar la Seguridad

Implementar marcos de reversibilidad y trazabilidad.

### 2. Alfabetización Digital

Capacitación del personal para interactuar con agentes.

### 3. Regulación Clara

Marcos éticos y operativos antes del despliegue.

### 4. Cultura Institucional

Responsabilidad final siempre en el humano.

NotebookLM

# Referencias Bibliográficas


- 1. Bandi A. The Rise of Agentic AI: A Review of Definitions and Architectures (2025).
- 2. Piccialli F. AgentAI: A comprehensive survey on autonomous agents (2025).
- 3. Joshi J. The Evolution of Agentic AI (2025).
- 4. Nisa U. Agentic AI: The age of reasoning – A review (2025).
- 5. Chaffer TJ. Decentralized Governance of Autonomous AI Agents (2024).
- 6. Pandey R. The Agentic AI Governance Framework (2025).
- 7. Papagiannidis E. Gobernanza responsable de la inteligencia artificial (2025).
- 8. Raza S, et al. TRiSM for Agentic AI (2025).
- 9. Khan R, et al. AGENTSAFE: A Unified Framework (2025).
- 10. Batool A. AI Governance: a systematic literature review (2025).

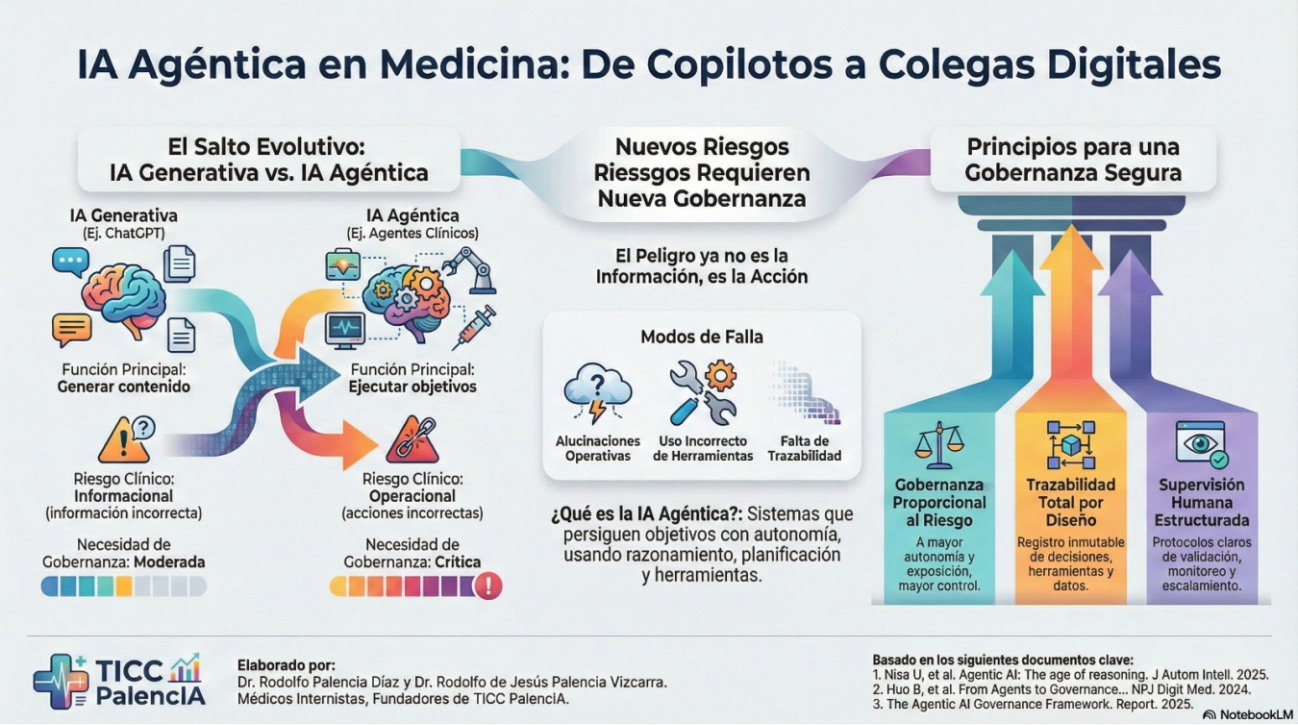

NotebookLM

# Evolución y gobernanza de la inteligencia artificial agénticas en educación y atención médica.

De “copilotos” a “colegas digitales”:la IA agéntica exige pasar de la fascinación tecnológica a una gobernanza clínica auditable, proporcional al riesgo y pedagógicamente responsable. [1] [3]

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra

Médicos Internistas
Universidad de Guadalajara
Instituto Mexicano del Seguro Social (IMSS)
Colegiados y Certificados (CMMI)
Fundadores del TICC
28 de enero de 2026




**Resumen**

La inteligencia artificial agéntica representa una evolución sustantiva respecto a los modelos generativos tradicionales, al incorporar capacidades de razonamiento iterativo, planificación, uso dinámico de herramientas, memoria y ejecución de acciones con distintos grados de autonomía. En educación médica y atención clínica, esta transición amplía de forma significativa las posibilidades de apoyo cognitivo, automatización de procesos, tutoría personalizada y soporte a la toma de decisiones; sin embargo, también incrementa la complejidad técnica, ética y organizacional, así como el riesgo potencial para la seguridad del paciente, la equidad educativa y la rendición de cuentas institucional.

Esta revisión analítica examina la evolución conceptual y técnica de la IA agéntica, sus principales patrones operativos y los modos de falla emergentes, destacando que el riesgo ya no reside únicamente en la generación de información incorrecta, sino en la capacidad del sistema para ejecutar acciones autónomas en entornos clínicos y educativos reales. A partir de la evidencia y marcos contemporáneos de gobernanza, se argumenta que los enfoques tradicionales de control de IA resultan insuficientes para sistemas agénticos, lo que obliga a adoptar modelos de gobernanza operacional, auditable y proporcional al riesgo, centrados en la trazabilidad, la supervisión humana estructurada y la gestión del ciclo de vida completo del sistema.

Asimismo, se discute la convergencia entre educación médica y práctica clínica, subrayando la necesidad de desarrollar competencias específicas en médicos, docentes e investigadores para evaluar, supervisar y gobernar sistemas de IA agéntica, más allá de su uso instrumental. Finalmente, se integran marcos metodológicos y de reporte reconocidos internacionalmente (PRISMA 2020, AMSTAR-II, GRADE, CONSORT-AI y GAMER) como pilares para una adopción responsable y científicamente sólida, concluyendo que la incorporación segura de la IA agéntica en salud depende menos del avance tecnológico aislado y más de la capacidad institucional para gobernarla de manera crítica, transparente y humanista.

**Introducción**

La IA agéntica (agentic AI) describe sistemas capaces de perseguir *objetivos con grados variables de autonomía*, ejecutando ciclos de *percepción-razonamiento-planificación-acción*, frecuentemente mediante uso dinámico de *herramientas, memoria, reflexión y colaboración multi-agente*.[1] Esta transición desde modelos generativos "conversasionales" hacia agentes con capacidad operativa incrementa el potencial de impacto en educación médica (p. ej., tutoría adaptativa, evaluación, apoyo administrativo) y en atención clínica (p. ej., soporte a decisión, estratificación de riesgo, automatización de flujos), pero también amplifica fallas críticas: alucinaciones con consecuencias operativas, acciones no autorizadas, deriva conductual, vulnerabilidades por "tool calling" y exposición regulatoria.[1] [4]

En salud, el problema central no es si la IA "acierta" en una respuesta aislada, sino si un sistema autónomo puede sostener seguridad, trazabilidad, rendición de cuentas y equidad a lo largo del ciclo de vida y en contextos reales.[3],[5] Esto obliga a integrar gobernanza específica para agentes (multipaso, selección dinámica de herramientas y coordinación entre agentes) que no queda completamente cubierta por marcos generales diseñados para IA tradicional.[3]

Métodos

Diseño. Revisión analítica basada principalmente en documentos compartidos por el solicitante (revisiones sobre IA agéntica, marcos de gobernanza agéntica y revisiones sobre gobernanza/responsabilidad de IA), complementada con guías metodológicas y de reporte internacionales para asegurar transparencia y rigor (PRISMA 2020, AMSTAR-2, GRADE, CONSORT-AI, GAMER).[6] [10]

Fuentes y elegibilidad. Se incluyeron: (1) una revisión sobre la evolución y patrones de IA agéntica (fases, patrones y entornos)[1]; (2) un marco operativo de gobernanza para sistemas agénticos con controles cuantificables (p. ej., trazabilidad, retención de evidencia, auditoría)[3]; (3) una revisión sistemática de literatura sobre gobernanza de IA con elementos aplicables a sistemas de salud (incluyendo gobernanza de IA clínica en sistemas grandes)[4]; (4) una síntesis sobre principios de gobernanza responsable (agencia humana, supervisión, auditoría, privacidad y justicia)[5]; y (5) un marco de competencias clínicas desde "agentes" hacia gobernanza en la era LLM, que vincula habilidades, práctica clínica y responsabilidad.[2]

Estrategia PRISMA 2020. Aunque esta revisión no pretende ser una revisión sistemática exhaustiva, se adoptaron componentes de PRISMA 2020 para documentar el flujo conceptual: identificación (documentos compartidos), selección (relevancia a IA agéntica + educación/atención + gobernanza), elegibilidad (contenido explícito sobre autonomía, riesgos, controles, o competencias), e inclusión final. [6]

Evaluación de calidad. Para documentos tipo revisión, se utilizaron dominios de AMSTAR-2 como lente crítica (claridad de pregunta, estrategia de búsqueda, justificación de exclusiones, evaluación de sesgo, y adecuación de síntesis).[7] Para recomendaciones prácticas (gobernanza y educación clínica), se utilizó GRADE como marco de "fuerza de recomendación" (fuerte vs condicional) y "certeza" (alta a muy baja), reconociendo que la evidencia en gobernanza suele ser predominantemente observacional, conceptual o de implementación.[8] Para investigación clínica con IA, se incorporó CONSORT-AI como estándar de reporte de ensayos con componente de IA.[9] Para transparencia en uso de herramientas de IA generativa en investigación y redacción, se incorporó GAMER.[10]

**Resultados**

1. Evolución: de sistemas reactivos a agentes multimodales colaborativos La revisión "Agentic AI: The reasoning" traza la evolución de la IA agéntica

en fases hasta la era actual de agentes multimodales y colaborativos, impulsada por aprendizaje por refuerzo, redes neuronales y LLMs.[1] En esta síntesis se destacan cinco patrones operativos que definen el comportamiento agéntico: uso de herramientas, reflexión, ReAct, planificación y colaboración multi-agente.[1]

Implicación para salud y educación: estos patrones convierten a la IA en un ejecutor de flujos (no solo un generador de texto). El salto de riesgo aparece cuando el sistema: (a) selecciona herramientas de forma dinámica,(b) mantiene memoria persistente (susceptible a "poisoning" o contaminación), y (c) actúa en sistemas clínicos o académicos reales (EHR/LMS, bases de datos, mensajería institucional ).[3,4]

2 Casos y modos de falla porque la gobernanza debe ser "operacional" y no solo "ética".
El marco de gobernanza agéntica compartido enfatiza que la adopción ha sobrepasado la definición de mecanismos de gobernanza y describe fallas plausibles y costosas cuando no existe supervisión humana ni trazabilidad: alucinaciones con datos sensibles o financieros, incumplimientos de privacidad multi-jurisdicción, decisiones discriminatorias y ausencia de auditabilidad.[3] Además, propone clasificar sistemas por nivel de autonomía y "superficie de exposición" (interno vs externo)destacando que no todo agente merece el mismo grado de control.[3]

En paralelo, la revisión sobre gobernanza responsable identifica "agencia y supervisión humana"como principio central , distinguiendo supervisión prospectiva (planificación), monitoreo continuo y análisis retrospectivo tras incidentes; subraya que la falta de auditabilidad y transparencia facilita la transferencia indebida de responsabilidad moral (culpar al sistema)[5]

3. Controles específicos para IA agéntica: del "cumplimiento declarativo" a métricas verificables.
El marco de gobernanza agéntica plantea que NIST AI RMF e ISO/IEC 42001 aportan fundamentos, pero no cubren de forma suficiente capacidades propias de agentes (decisión autónoma multi-paso y selección dinámica de herramientas).[3] Por ello propone controles añadidos: orquestación multi-agente y asignación de responsabilidad; gobernanza de tool-calling; reglas dinámicas de escalamiento; cumplimiento multi- jurisdicción; y auditabilidad del razonamiento.[3]

Operativamente, resalta medidas como validación pre-despliegue (no solo happy paths) monitoreo/registro inmutable, y capacidad de reconstrucción post- noc; por ejemplo, "equipo de registro cada decisión con contexto y mantener trazabilidad con evidencias a prueba de manipulación.[3] Este punto es crítico en salud y educación, donde auditoría y explicabilidad no son "lujos" sino requisitos para seguridad del paciente, evaluación justa y rendición de cuentas institucional.[2,3,5]

4. Convergencia educación clínica: competencias y gobernanza como binomio.
El artículo sobre "De los agentes a la gobernanza" enfatiza habilidades esenciales para clínicos en la era LLM, conectando la progresión desde uso instrumental hacia comprensión de riesgos, validación, monitoreo y gobernanza aplicada.[2] En términos prácticos: a medida que un sistema pasa de tutor conversacional a agente que ejecuta tareas (p. ej., generar materiales de curso, calificar, registrar incidencias, proponer planes clínicos o disparar órdenes), el dominio requerido del usuario y de la institución cambia: se requieren competencias para auditar, establecer límites, interpretar incertidumbre y operar protocolos de escalamiento.[2,3]

**Discusión**

1. Propuesta de modelo integrado para salud y educación: "gobernanza proporcional al riesgo"
A partir de los documentos analizados, la gobernanza efectiva de IA agéntica en medicina y educación se sostiene en cinco tesis:

**Tesis 1.** La unidad de análisis ya no es el "modelo", sino el "sistema agéntico" (modelo + memoria + herramientas + permisos + datos + orquestación +humanos).[1,3]
Implicación: la evaluación debe incluir rutas de ejecución, permisos, y puntos de falla por tool-calling, no solo desempeño en benchmarks.

**Tesis 2.** La autonomía debe estar acoplada a exposición y daño potencial; el control debe ser proporcional (no uniforme).[3]

Aplicación clínica: un agente interno que resume bibliografía para docentes no requiere los mismos controles que un agente que sugiere ajustes terapéuticos o interactúa con pacientes.

**Tesis 3.** La trazabilidad deber ser un requisito de diseño (evidence-by-default) no una auditoría posterior.[3,5]
En salud, esto se traduce en bitácoras con contexto, versiones de modelo, fuentes consultadas, prompts relevantes, herramientas usadas y criterio de escalamiento a humano.

**Tesis 4.** La "agencia humana" se operacionaliza en tres momentos antes (validación y límites), durante (monitoreo y escalamiento), después (análisis de incidentes y aprendizaje institucional).[5]
Esto es coherente con una cultura clínica de seguridad: reporte, análisis de causa raíz, acciones correctivas y prevención.

**Tesis 5.** Educación médica y práctica clínica comparten la misma amenaza: automatizar sin gobernar degrada confianza, equidad y calidad.[2,3,5]
La institución debe tratar el LMS/EHR como entornos regulados: control de acceso, minimización de datos, y evaluación continua.

2. Recomendaciones (con lente GRADE) para implementación responsable Dado que parte de la evidencia es principalmente conceptual/observacional (certeza típicamente baja a moderada), las recomendaciones se formulan como políticas prudenciales, alineadas con prevención de daño:

**Recomendación 1** (Fuerte; certeza moderada): Inventariar y clasificar todo sistema agéntico por autonomía y superficie de exposición, y asignar controles por nivel de riesgo.[3,8]
Justificación: la matriz de riesgo y la lógica de proporcionalidad permiten priorizar gobernanza donde el daño potencial es mayor.[3]

**Recomendación 2** (Fuerte; certeza moderada): Implementar trazabilidad y registro inmutable de "extremo a extremo" (decisiones, herramientas, contexto versionado), con capacidad de reconstrucción post-hoc.[3,5,8]
Justificación: sin evidencia reconstruible, no existe rendición de cuentas ni aprendizaje tras incidentes, especialmente en entornos regulados.[3,5]

**Recomendación 3** (Fuerte; certeza baja-moderada): Establecer protocolos explícitos de human-in-the-loop con umbrales de escalamiento, allow-lists de herramientas, y límites de permisos (principio de mínimo privilegio).[3,5,8]
Justificación: tool-calling y acciones autónomas son los multiplicadores de riesgo más relevantes para daño operativo y seguridad.[3]

**Recomendación 4** (Condicional; certeza baja): Desarrollar competencias clínicas y docentes centradas en validación, monitoreo, sesgos, privacidad y gobernanza aplicada (no solo promoting).[2,8]
Justificación: el artículo de competencias vincula el uso seguro con habilidades de evaluación y gobernanza; sin alfabetización operativa, la institución delega riesgos al usuario final.[2]

**Recomendación 5** (Fuerte; certeza alta para reporte): Para investigación y publicación, adherirse a PRISMA 2020 (síntesis), AMSTAR-2 (calidad de revisiones), CONSORT-AI (ensayos con IA) y GAMER (transparencia en uso de GenAI).[6,7,9,10]
Justificación: son estándares de reporte/método con amplia adopción y reducen opacidad, sesgos de reporte y ambigüedad sobre el rol de IA en la evidencia.

3. Implicaciones para investigación y evaluación en salud
   La literatura revisada sobre gobernanza en salud sugiere que la implementación segura requiere elementos organizacionales (política, rendición de cuentas, gobernanza de datos) además de componentes técnicos.[4] Esto converge con el principio de auditoría/controles y con la necesidad de supervisión humana estructurada.[3,5] Para ensayos clínicos o intervenciones con componente IA, CONSORT-AI aporta el estándar mínimo para reportar adecuadamente qué hace la IA, cómo se integra al flujo clínico, y cómo se evalúa su desempeño y seguridad.[9]

**Conclusiones**

La IA agéntica representa un cambio de fase: de sistemas que "responden" a sistemas que "operan".[1] En educación médica y atención clínica esto crea oportunidades reales de mejora de productividad, personalización y soporte cognitivo, pero también nuevos vectores de daño: autonomía multi-paso, tool-calling, memoria

persistente y coordinación multi-agente.[1] [3] La respuesta responsable no es frenar la adopción, sino institucionalizar gobernanza operacional, auditable y proporcional al riesgo, alineada con supervisión humana y trazabilidad.[3],[5]

En paralelo, la educación médica debe evolucionar: formar clínicos y docentes no solo en uso, sino en verificación, límites, sesgos, privacidad y gobernanza aplicada.[2] Finalmente, la producción científica y la evaluación de evidencia deben sostenerse en estándares robustos (PRISMA 2020, AMSTAR-2, GRADE, CONSORT-IA, GAMER) para evitar que la "eficiencia" introduzca opacidad metodolóica.[6] [10]

**Tabla 1. Evolución de la IA en salud y educación: de modelos reactivos a IA agéntica**

| Etapa evolutiva | Características principales | Nivel de autonomía | Riesgo operativo | Ejemplos en salud / educación |
|---|---|---|---|---|
| **IA tradicional** | Algoritmos deterministas, reglas fijas | Nulo | Bajo | Scores clínicos clásicos |
| **Machine Learning** | Predicción basada en datos históricos | Bajo | Bajo–moderado | Predicción de riesgo CV |
| **Deep Learning** | Representaciones profundas, menor explicabilidad | Bajo–moderado | Moderado | Imagenología, NLP clínico |
| **IA generativa (LLM)** | Generación de texto/contenido, razonamiento probabilístico | Moderado | Moderado | Resúmenes clínicos, tutor virtual |
| **IA agéntica** | Planificación, uso de herramientas, memoria, ejecución | **Alto** | **Alto** | Agentes clínicos, tutores autónomos |

**Tabla 2. Comparación entre IA generativa y IA agéntica**

| Dimensión | IA generativa | IA agéntica |
|---|---|---|
| **Función principal** | Generar contenido | Ejecutar objetivos |
| **Ciclo de acción** | Respuesta única | Percepción–razonamiento–acción |
| **Uso de herramientas** | Limitado | Dinámico (tool calling) |
| **Memoria** | Contextual, corta | Persistente |
| **Capacidad de acción** | Pasiva | Activa |
| **Riesgo clínico** | Informacional | Operacional |
| **Necesidad de gobernanza** | Moderada | Crítica |

**Tabla 3. Niveles de autonomía y gobernanza proporcional al riesgo**

| Nivel de autonomía | Descripción | Ejemplo | Gobernanza requerida |
|---|---|---|---|
| **Nivel 0** | Sin autonomía | Calculadora médica | Validación estándar |
| **Nivel 1** | Asistencia | Resumen de guías | Revisión humana |
| **Nivel 2** | Recomendación | Sugerencia diagnóstica | HITL obligatorio |
| **Nivel 3** | Acción supervisada | Generación de órdenes sugeridas | Auditoría + escalamiento |
| **Nivel 4** | Acción autónoma | Agente clínico operativo | Gobernanza completa, trazabilidad total |

**Tabla 4. Gobernanza tradicional vs gobernanza agéntica**

| Dimensión | Gobernanza tradicional de IA | Gobernanza de IA agéntica |
|---|---|---|
| **Objeto de control** | Modelo | Sistema completo |
| **Enfoque** | Ético–normativo | Operacional–auditivo |
| **Trazabilidad** | Parcial | End-to-end |
| **Supervisión humana** | Implícita | Estructurada |
| **Gestión del riesgo** | Estática | Dinámica |
| **Auditoría** | Posterior | Continua |

**Tabla 5. Modos de falla en IA agéntica y mitigación**

| Modo de falla | Descripción | Riesgo clínico/educativo | Medida de mitigación |
|---|---|---|---|
| **Alucinación operativa** | Información falsa que guía acción | Alto | HITL + verificación |
| **Tool misuse** | Uso incorrecto de herramientas | Alto | Allow-list y permisos |
| **Deriva conductual** | Cambio progresivo del comportamiento | Moderado | Monitoreo continuo |
| **Sesgo acumulativo** | Memoria persistente sesgada | Alto | Reinicio y auditoría |
| **Falta de trazabilidad** | Imposible reconstruir decisiones | Crítico | Logging inmutable |

**Tabla 6. Roles humanos en sistemas de IA agéntica**

| Rol | Función | Momento de intervención |
|---|---|---|
| **Diseñador** | Define límites y arquitectura | Pre-despliegue |
| **Clínico/docente** | Valida resultados | Uso activo |
| **Supervisor** | Monitorea desempeño | Tiempo real |
| **Auditor** | Analiza incidentes | Post-evento |
| **Comité institucional** | Gobernanza y políticas | Ciclo completo |

**Tabla 7. Educación médica: competencias según nivel de IA**

| Nivel tecnológico | Competencia requerida |
|---|---|
| **IA básica** | Alfabetización digital |
| **IA generativa** | Evaluación crítica de contenido |
| **IA agéntica** | Gobernanza, auditoría, ética aplicada |
| **Multi-agente** | Gestión del riesgo sistémico |

**Tabla 8. Marcos metodológicos y su aplicación**

| Marco | Tipo | Aplicación principal |
|---|---|---|
| PRISMA 2020 | Reporte | Revisiones sistemáticas |
| AMSTAR-II | Calidad | Evaluación de revisiones |
| GRADE | Recomendaciones | Fuerza y certeza |
| CONSORT-AI | Ensayos | Intervenciones con IA |
| GAMER | Transparencia | Uso de IA generativa |

**Tabla 9. Educación vs atención clínica: convergencias y diferencias**

| Dimensión | Educación médica | Atención clínica |
|---|---|---|
| **Riesgo principal** | Evaluación injusta | Daño al paciente |
| **Entorno** | LMS | EHR |
| **Gobernanza** | Académica | Clínica–legal |
| **Supervisión** | Docente | Médico responsable |
| **Estándares** | PRISMA, GAMER | CONSORT-AI, GRADE |

**Tabla 10. Síntesis operativa para instituciones de salud**

| Elemento | Requisito mínimo |
|---|---|
| **Inventario de IA** | Clasificación por riesgo |
| **Políticas** | Autonomía y límites claros |
| **Trazabilidad** | Registros auditables |
| **Capacitación** | Competencias en gobernanza |
| **Evaluación continua** | KPIs y revisión periódica |

**Bibliografía**


1. Nisa U, Shirazi M, Saip MA, et al. Agentic AI: The age of reasoning-A. review. J Autom Intell. 2025;xxx(xxx):xxx. doi:10.1016/j.jai.2025.08.003.
2. Huo B, Fridsma DB, et al. From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era. NPJ Digit Med. 2024. doi:10.1038/s41746-024-01263-7.
3. The Agentic AI Governance Framework: Operational Controls for Autonomous and Multi-Agent Systems. Report. 2025. (No DOI disponible en el documento compartido).
4. AI Governance: A Systematic Literature Review. 2024. (Documento compartido; incluye gobernanza de IA clínica en sistemas de salud).
5. Papagiannidis E, et al. Responsible artificial intelligence governance. J Strateg Inf Syst. 2025;34:101885. doi:10.1016/j.jsis.2025.101885.
6. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
7. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008. doi:10.1136/bmj.j4008.
8. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924. doi:10.1136/bmj.39489.470347.AD.
9. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364 1374. doi:10.1038/s41591-020-1034-x.
10. Luo X, et al. Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER Statement. BMJ Evid Based Med. 2025. doi:10.1136/bmjebm-2024-113323.


**Glosario de términos**

Inteligencia Artificial (IA)
Campo de la informática dedicado al desarrollo de sistemas capaces derealizar tareas que normalmente requieren inteligencia humana, como razonamiento, aprendizaje, toma de decisiones y reconocimiento de patrones.

IA generativa (Generative AI)
Subconjunto de la IA que utiliza modelos entrenados con grandes volúmenes de datos para generar contenido nuevo (texto, imágenes, código, audio), manteniendo coherencia estadística con los datos de entrenamiento.

Modelo de Lenguaje de Gran Escala (LLM, Large Language Model) Modelo de IA entrenado con grandes corpus textuales para predecir la siguiente palabra o token, permitiendo tareas como redacción, resumen, traducción y razonamiento probabilístico.

IA agéntica (Agentic AI)
Arquitectura de IA en la que un sistema puede perseguir objetivos de manera semi-autónoma o autónoma mediante ciclos iterativos de percepción, razonamiento, planificación y acción, incluyendo uso dinámico de herramientas, memoria y colaboración con otros agentes.

Agente de IA
Entidad computacional que percibe su entorno, toma decisiones y ejecuta acciones orientadas a objetivos definidos, con distintos niveles de autonomía y supervisión humana.

Multi-agente
Sistema compuesto por múltiples agentes que interactúan y coordinan acciones, compartiendo o no memoria y objetivos, lo que incrementa complejidad, eficiencia potencial y riesgo sistémico.

Autonomía
Grado en que un sistema de IA puede tomar decisiones y ejecutar acciones sin intervención humana directa. En salud, la autonomía debe ser siempre proporcional al riesgo clínico.

Human-in-the-Loop (HITL)
Modelo de supervisión en el que un humano valida, corrige o autoriza decisiones del sistema de IA antes, durante o después de su ejecución.

Human-on-the-Loop
Modalidad de supervisión en la que el humano no interviene de forma continua, pero puede detener o corregir el sistema si se detectan desviaciones.

Tool calling (uso de herramientas)
Capacidad de un agente de IA para invocar herramientas externas (bases de datos, APIs, sistemas clínicos, calculadoras) como parte de su razonamiento y ejecución.

Memoria persistente
Almacenamiento de información entre interacciones que permite al agente datos, privacidad y sesgos acumulativos.
Alucinación
Generación de información falsa, no verificable o incorrecta presentada como verdadera por un modelo de IA. En sistemas agénticos, puede traducirse en acciones erróneas.

Razonamiento en cadena (Chain-of-Thought)
Proceso interno mediante el cual un modelo descompone un problema en pasos intermedios. No siempre visible ni auditable para el usuario final.

ReAct (Reasoning and Acting)
Paradigma en el que el modelo alterna explícitamente entre razonamiento y acción, especialmente relevante en sistemas agénticos con uso de herramientas.

Gobernanza de IA
Conjunto de políticas, procesos, roles y controles técnicos y organizacionales destinados a asegurar que los sistemas de IA sean seguros, éticos, auditables y alineados con objetivos institucionales.

Gobernanza agéntica
Extensión de la gobernanza de IA tradicional, enfocada específicamente en sistemas con autonomía operativa, múltiples pasos de decisión y capacidad de acción directa.

Proporcionalidad al riesgo
Principio según el cual el nivel de control, supervisión y validación de un sistema de IA debe corresponder al daño potencial que puede causar.

Auditabilidad
Capacidad de reconstruir y revisar el comportamiento de un sistema de IA, incluyendo decisiones, datos utilizados, herramientas invocadas y resultados generados.

Trazabilidad
Registro continuo y verificable del flujo de decisiones y acciones de un sistema de IA a lo largo de su ciclo de vida.

Transparencia algorítmica
Grado en que el funcionamiento, limitaciones y supuestos de un sistema de IA
son comprensibles para usuarios y auditores.

Sesgo algorítmico
Distorsión sistemática en los resultados de un sistema de IA que afecta de manera injusta a ciertos grupos, frecuentemente originada en los datos de entrenamiento o diseño.

Privacidad de datos
Protección de la información personal y sensible frente a accesos no autorizados, uso indebido o filtraciones, especialmente crítica en salud y educación.

Ciclo de vida del sistema de IA
Conjunto de etapas que abarcan diseño, desarrollo, validación, despliegue, monitoreo, actualización y retiro del sistema.

Supervisión prospectiva, concurrente y retrospectiva
Enfoque de control que contempla: validación previa al despliegue, monitoreo en tiempo real y análisis posterior a incidentes.

Curaduría del conocimiento
Rol del médico o docente como evaluador crítico de la información generada por IA, integrándola con evidencia científica y juicio profesional.

**Glosario de abreviaturas**

IA - Inteligencia Artificial
Iag - Inteligencia Artificial agéntica
LLM - Large Language Model
HITL - Human-in-the-Loop
EHR - Electronic Health Record (Expediente Clínico Electrónico)
LMS - Learning Management System
API - Application Programming Interface
PRISMA 2020 - Preferred Reporting Items for Systematic Reviews and Meta- Analyses
AMSTAR-II - A Measurement Tool to Assess Systematic Reviews
GRADE - Grading of Recommendations Assessment, Development and Evaluation
CONSORT-AI - Consolidated Standards of Reporting Trials Artificial Intelligence
GAMER - Reporting guideline for the use of Generative Artificial Intelligence tools in Medical Research
NIST AI RMF - National Institute of Standards and Technology Artificial Intelligence Risk Management Framework
ISO/IEC 42001 - Norma internacional de sistemas de gestión para IA KPI - Key Performance Indicator

# Evolución y gobernanza de la inteligencia artificial agéntica en educación y atención médica

De la fascinación tecnológica a una gobernanza clínica auditable, proporcional al riesgo y pedagógicamente responsable.

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de Jesús Palencia Vizcarra
Médicos Internistas

TICC PalencIA

NotebookLM

## De 'Copilotos' a 'Colegas Digitales': El Ciclo Agéntico

La IA agéntica representa una evolución sustantiva respecto a los modelos generativos. Ya no se trata solo de generar texto, sino de perseguir objetivos con autonomía [1].

La **IA se convierte en un ejecutor de flujos**, no solo en un generador de contenido.


NotebookLM

# Anatomía Comparada: IA Generativa vs. IA Agéntica

| Dimensión | IA Generativa (Tradicional) | IA Agéntica (Nueva Era) |
|---|---|---|
| **Función Principal** | Generar contenido (Texto/Imagen) | Ejecutar objetivos complejos |
| **Ciclo Operativo** | Respuesta única (Input -> Output) | Ciclo continuo (Percepción -> Acción) |
| **Herramientas** | Limitado / Ninguna | Dinámico (Tool Calling) |
| **Memoria** | Ventana de contexto corta | Persistente y reflexiva |
| **Riesgo Clínico** | Informacional (Dato falso) | Operacional (Acción errónea) [1, 3] |

[1], [3]

NotebookLM

Clinical Consultancy

# Nuevos Vectores de Falla: Del Error de Sintaxis al Daño Operativo

En salud, el problema central ya no es solo si la IA "acierta", sino las consecuencias de sus acciones autónomas.

### Alucinación Operativa

Información falsa que guía una acción clínica real.

### Tool Misuse

Uso incorrecto de herramientas digitales (ej. acceso no autorizado).


### Deriva Conductual

Cambio progresivo del comportamiento sin supervisión.

### Falla de Trazabilidad

Imposibilidad de reconstruir la decisión (Caja negra) [3, 4].

**Riesgo:** Transferencia indebida de responsabilidad moral ('culpar al sistema').

NotebookLM

# Metodología de la Revisión: Rigor Científico

Esta revisión analítica integra marcos internacionales para asegurar transparencia y rigor [6-10].

**PRISMA 2020**

Flujo conceptual y selección sistemática de documentos [6].

**AMSTAR-2**

Evaluación crítica de la calidad de las revisiones incluidas [7].

**GRADE**

Determinación de la fuerza de las recomendaciones [8].

**CONSORT-AI / GAMER**

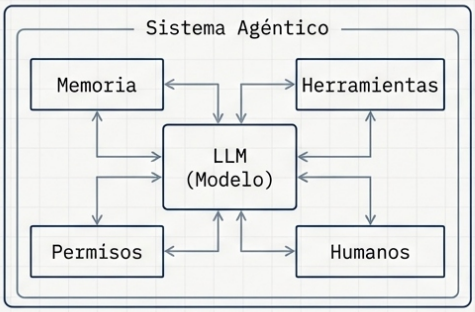
Estándares de reporte y transparencia en investigación [9, 10].

**La adopción tecnológica debe seguir la misma disciplina metodológica que la práctica clínica.**

NotebookLM

# Las 5 Tesis de la Gobernanza Agéntica (Parte I)

## Tesis 1: El Sistema sobre el Modelo




La unidad de análisis ya no es el "modelo" aislado, sino el sistema agéntico completo. Evaluar solo el chat es insuficiente; se debe auditar el flujo operativo.

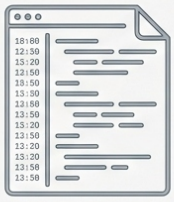
## Tesis 2: Proporcionalidad del Riesgo


El control debe ser proporcional al daño potencial. Un agente de consulta bibliográfica no requiere la misma gobernanza que uno que ejecuta órdenes clínicas [3].

NotebookLM

# Las 5 Tesis de la Gobernanza Agéntica (Parte II)

### Tesis 3: Trazabilidad por Diseño



'Evidence by default'. Sin registros inmutables de decisiones y herramientas usadas, no existe rendición de cuentas profesional.

### Tesis 4: Agencia Humana Estructurada


La supervisión humana se operacionaliza en momentos específicos, no como un concepto abstracto.

### Tesis 5: Convergencia Educativa-Clínica

Educación y práctica comparten la misma amenaza: automatizar sin gobernar degrada la confianza y la seguridad del paciente [2, 3].

NotebookLM

# De la Ética Normativa al Control Operacional

Por qué la gobernanza debe ser "operacional" y no solo "ética".

| | Gobernanza Tradicional | Gobernanza Agéntica |
|---|---|---|
| **Enfoque** | Ético / Declarativo | Operacional / **Auditable** |
| **Objeto** | Modelo (LLM) | Sistema Completo |
| **Trazabilidad** | Parcial | **End-to-End** (Logs inmutables) |
| **Supervisión** | Implícita | Estructurada (HITL) |

***"Pasar del cumplimiento declarativo a una gobernanza clínica auditable."***

NotebookLM

# Espectro de Autonomía y Supervisión Requerida


NotebookLM

# El Rol del Médico: Human-in-the-Loop (HITL)

La supervisión humana como componente de seguridad del paciente, no como obstáculo.

**Supervisor**
Monitorea el desempeño en tiempo real (HITL).

**Clínico / Docente**
Valida resultados y utiliza el sistema activamente.

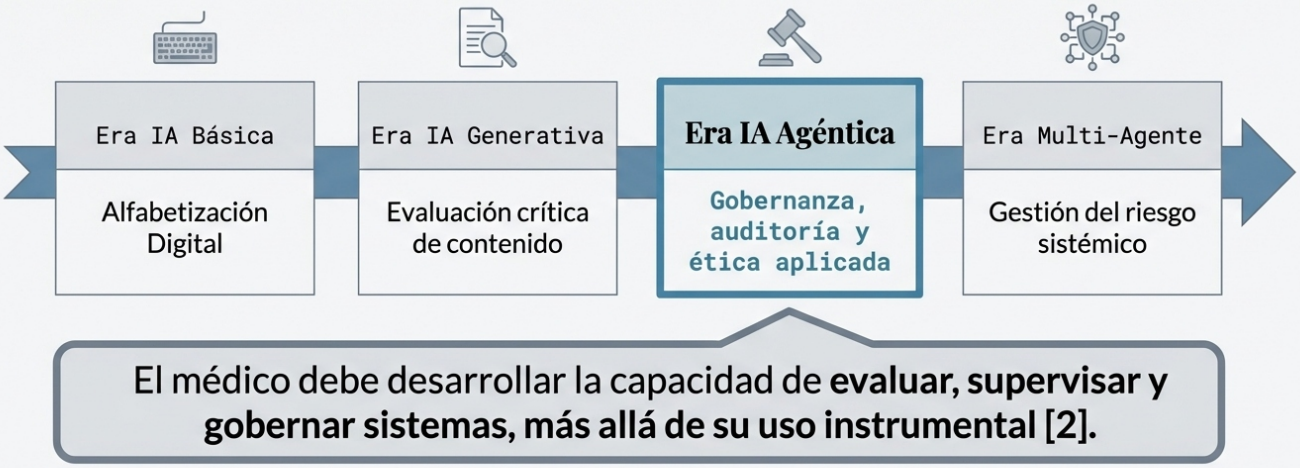
`System`
`Datos de Entrada | Algoritmo IA | Resultados Clínicos`

**Auditor**
Analiza incidentes y bitácoras post-evento (Análisis de Causa Raíz) [5].

**Riesgo a mitigar:** La falta de **auditabilidad** facilita la transferencia indebida de responsabilidad moral ('**culpar al sistema**').

NotebookLM

## Nuevas Competencias Clínicas en la Era Agéntica



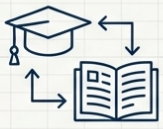

El médico debe desarrollar la capacidad de **evaluar, supervisar y gobernar sistemas, más allá de su uso instrumental [2].**



## Convergencia: Educación Médica y Práctica Clínica

Propuesta de modelo integrado para salud y educación: "gobernanza del binomio".

### Entorno Educativo (LMS)



**Riesgo Principal:** Evaluación injusta y equidad.

**Necesidad:** Gobernanza Académica.

### Amenazas Compartidas

Desarrollar habilidades de gobernanza en la residencia prepara para la práctica clínica.

### Entorno Clínico (EHR)

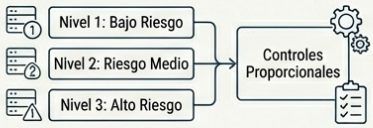
**Riesgo Principal:** Daño al paciente y seguridad.

**Necesidad:** Gobernanza Legal/Clínica.

# Recomendaciones para la Implementación (GRADE)

**1** FUERTE **Inventariar y Clasificar**

Clasificar todo sistema por nivel de riesgo y autonomía. Asignar controles proporcionales [3, 8].




**2** FUERTE **Trazabilidad por Diseño**

Implementar registros auditables inmutables y capacidad de reconstrucción post-hoc [3, 5].


**3** FUERTE **Protocolos HITL**

Establecer umbrales explícitos de escalamiento a humanos y 'allow-lists' de herramientas.


**4** CONDICIONAL **Capacitación Continua**

Desarrollar competencias de validación y gobernanza en todo el personal [2].


NotebookLM

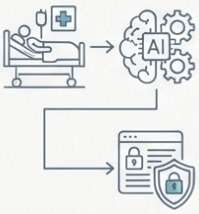
# Estándares de Reporte y Evidencia Científica

Evitando la opacidad en la investigación médica con IA.

## CONSORT-AI

**Estándar para ensayos clínicos.**

Define cómo reportar la intervención de IA, su integración en el flujo clínico y el análisis de seguridad [9].




## GAMER

**Estándar para redacción científica.**

Transparencia en el uso de IA generativa para la creación de manuscritos y síntesis de evidencia [10].


**Objetivo:** Asegurar que la evidencia clínica sea reproducible, transparente y válida.

NotebookLM

# Referencias Bibliográficas


1. Nisa U, Shirazi M, Saip MA, et al. Agentic AI: The age of reasoning – A review. J Autom Intell. 2025;xxx(xxx):xxx.
2. Huo B, Fridsma DB, et al. From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era. NPJ Digit Med. 2024.
3. The Agentic AI Governance Framework: Operational Controls for Autonomous and Multi-Agent Systems. Report. 2025.
4. AI Governance: A Systematic Literature Review. 2024.
5. Papagiannidis E, et al. Responsible artificial intelligence governance. J Strateg Inf Syst. 2025;34:101885.
6. Page MJ, et al. The PRISMA 2020 statement. BMJ. 2021;372:n71.
7. Shea BJ, et al. AMSTAR 2: a critical appraisal tool. BMJ. 2017;358:j4008.
8. Guyatt GH, et al. GRADE: an emerging consensus. BMJ. 2008;336:924.
9. Liu X, et al. Reporting guidelines for clinical trial reports … CONSORT-AI extension. Nat Med. 2020;26:1364–1374.


TICC PalencIA

NotebookLM

# Agentes de IA y Autonomía Operativa: Análisis Crítico de Implicaciones Éticas, Laborales y de Gobernanza.

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra

Médicos Internistas
Universidad de Guadalajara
Instituto Mexicano del Seguro Social (IMSS)
Colegiados y Certificados (CMMI)
Fundadores del TICC
28 de enero de 2026

**Resumen**

La evolución de sistemas de inteligencia artificial desde herramientas asistenciales hacia agentes autónomos capaces de ejecutar tareas complejas representa un cambio paradigmático en la interacción humano-máquina. Esta revisión analítica examina críticamente las dimensiones éticas, laborales, de responsabilidad y gobernanza asociadas con los agentes de IA, con énfasis en modelos recientes como Claude Cowork y capacidades agénticas emergentes. Se analizan marcos regulatorios internacionales, evidencia empírica sobre automatización, modelos de responsabilidad legal y consideraciones éticas fundamentales para el desarrollo responsable de estas tecnologías.

**Introducción**

Los sistemas de inteligencia artificial han transitado desde funciones consultivas hacia capacidades operativas con grados crecientes de autonomía (1). Francisco Carvajal plantea preguntas fundamentales sobre esta transición: ¿quién asume responsabilidad cuando un agente de IA toma decisiones erróneas?, ¿estamos automatizando tareas o delegando criterio profesional?, y ¿constituye esto una evolución inevitable o una cesión prematura de autonomía? (2).

Los plugins agénticos, como Claude Cowork de Anthropic, representan sistemas capaces de ejecutar flujos de trabajo completos, interactuar con documentos empresariales y tomar decisiones operativas con supervisión humana variable (3). Esta capacidad plantea desafíos inéditos en materia de responsabilidad, transparencia algorítmica y redistribución del trabajo cognitivo.

**Evolución Tecnológica: De Asistentes a Agentes**

Taxonomía de Sistemas de IA

Russell y Norvig establecen una distinción fundamental entre agentes reactivos, deliberativos y adaptativos (4). Los sistemas actuales incorporan:

1. Agentes conversacionales: Limitados a interacciones lingüísticas sin capacidad ejecutiva
2. Copilotos: Sugieren acciones que requieren validación humana explícita
3. Agentes autónomos: Ejecutan secuencias de tareas con supervisión mínima

La Unión Europea clasifica estos sistemas según niveles de riesgo, considerando agentes con capacidad decisional autónoma como "sistemas de alto riesgo" que requieren evaluaciones de conformidad previas a su despliegue (5).

Capacidades Técnicas Emergentes

Los large language models (LLMs) con arquitectura transformer han demostrado capacidades emergentes de razonamiento, planificación y uso de herramientas (tool use) (6). Anthropic documenta que Claude 3.5 y Claude 4 pueden:

- Descomponer objetivos complejos en subtareas ejecutables
- Interactuar con APIs y sistemas empresariales
- Mantener contexto prolongado (ventanas de hasta 200,000 tokens)
- Autoevaluar calidad de outputs y solicitar aclaraciones (7)

Estas capacidades transforman la IA de "oráculo consultivo" a "ejecutor operativo", modificando fundamentalmente la división del trabajo cognitivo.

Dimensión Ética: Autonomía, Responsabilidad y Transparencia

Marcos Éticos en IA

Los Principios de Asilomar para IA Beneficiosa establecen que sistemas autónomos deben incorporar (8):

1. Alineación de valores: Coherencia entre objetivos del sistema y valores humanos
2. Transparencia: Explicabilidad de procesos decisionales
3. Responsabilidad: Cadenas claras de responsabilidad
4. Dignidad humana: Preservación de agencia y autodeterminación

Floridi y Cowls proponen un marco de "ética de la información" que enfatiza la distribución equitativa de beneficios, prevención de daños, respeto a autonomía humana y justicia procedimental (9).

Dilemas Éticos Específicos

Delegación vs. Automatización

Carvajal distingue entre automatizar tareas mecánicas y delegar criterio profesional. La literatura en ética médica diferencia entre:

- Automatización apropiada: Tareas algorítmicas sin componente valorativo significativo
- Delegación problemática: Decisiones que requieren juicio contextual, consideración de valores y responsabilidad moral (10)

Coeckelbergh argumenta que delegar decisiones moralmente significativas a sistemas de IA constituye una abdicación de responsabilidad, no una transferencia legítima (11).

El Problema de la Caja Negra

Los modelos de lenguaje basados en redes neuronales profundas operan como "cajas negras" donde no es posible rastrear cadenas causales específicas entre inputs y outputs (12). Esto genera:

1. Opacidad epistémica: Imposibilidad de justificar decisiones mediante razonamiento explícito
2. Riesgo de sesgo oculto: Perpetuación de discriminaciones sin mecanismos de detección
3. Erosión de confianza: Dificultad para establecer fiabilidad sin comprensión del proceso

La Ley de IA de la Unión Europea (AI Act) exige documentación técnica exhaustiva y explicaciones comprensibles para sistemas de alto riesgo (13).

Poder y Control

Winner analiza cómo artefactos tecnológicos incorporan relaciones de poder, argumentando que sistemas autónomos pueden concentrar autoridad decisional sin mecanismos democráticos de control (14). En contextos laborales, esto plantea:

- Asimetría informacional: Trabajadores sujetos a decisiones algorítmicas sin transparencia
- Deskilling: Erosión de competencias profesionales por dependencia tecnológica
- Precariedad: Reducción de autonomía laboral y subordinación a sistemas opacos

Dimensión Laboral: Transformación del Trabajo Cognitivo

Evidencia Empírica sobre Automatización

Frey y Osborne estimaron en 2013 que 47% de empleos en Estados Unidos tenían alta probabilidad de automatización (15). Estudios posteriores refinaron estas proyecciones:

- Arntz et al. (2016): 9% de empleos con riesgo real alto al considerar tareas específicas, no ocupaciones completas (16)
- McKinsey Global Institute (2023): 30% de horas trabajadas en Estados Unidos podrían automatizarse para 2030 con tecnologías actuales (17)
- OCDE (2023): Énfasis en "transformación de empleos" más que "reemplazo", con 27% de trabajadores en empleos de alto riesgo (18)

Efectos Diferenciados por Sector

Acemoglu y Restrepo documentan heterogeneidad sectorial significativa (19):

1. Trabajo rutinario cognitivo: Alta susceptibilidad (contabilidad, análisis de datos básico, redacción estandarizada)

2. Trabajo manual no rutinario: Baja susceptibilidad (oficios especializados, cuidados personales)
3. Trabajo cognitivo complejo: Transformación más que reemplazo (medicina, derecho, ingeniería)

Brynjolfsson et al. identifican "complementariedad aumentada" donde IA incrementa productividad de trabajadores cualificados sin sustituirlos completamente (20).

Implicaciones para Profesiones Cognitivas

En medicina, estudios documentan:

- Diagnóstico por imagen: Algoritmos igualan o superan radiólogos en detección de patología específica, pero integración clínica requiere expertise humano (21)
- Medicina de precisión: IA identifica patrones en datos genómicos, pero decisiones terapéuticas demandan consideración holística del paciente (22)
- Documentación clínica: Sistemas de reconocimiento de voz y resumen automático reducen carga administrativa, pero supervisión médica sigue siendo esencial (23)

Cabitza et al. advierten sobre "automation bias" donde médicos aceptan recomendaciones algorítmicas acríticamente, comprometiendo seguridad del paciente (24).

Responsabilidad Legal

Vacíos en Marcos Jurídicos Actuales

La responsabilidad legal tradicional asume agentes humanos con intencionalidad y capacidad de rendir cuentas. Los sistemas de IA desafían estos supuestos:

Modelo de Responsabilidad del Fabricante

Bajo derecho de productos defectuosos, fabricantes responden por daños causados por fallas de diseño, manufactura o advertencias inadecuadas (25). Aplicado a IA:

- Ventajas: Incentiva calidad y seguridad en desarrollo
- Limitaciones: Difícil probar causalidad en sistemas adaptativos; empresas pueden argumentar "uso indebido" por usuarios

Modelo de Responsabilidad del Usuario

Usuarios finales asumen responsabilidad por decisiones tomadas con asistencia de IA (26). Problemas:

- Asimetría de conocimiento: Usuarios no comprenden completamente funcionamiento del sistema
- Dilución de responsabilidad: Si IA ejecuta decisión, ¿usuario es realmente agente causal?

Modelo de Responsabilidad Distribuida

Propuesto por Matthias (2004), sugiere responsabilidad compartida entre desarrolladores, implementadores y usuarios según contribución causal (27). Requiere:

1. Trazabilidad de decisiones algorítmicas
2. Documentación exhaustiva de capacidades y limitaciones
3. Mecanismos de supervisión humana efectiva

Marco Regulatorio Europeo: AI Act

La Ley de Inteligencia Artificial de la UE (aprobada 2024, implementación progresiva hasta 2027) establece (28):

Clasificación por Riesgo

- Riesgo inaceptable: Prohibidos (manipulación subliminal, explotación de vulnerabilidades)
- Alto riesgo: Requisitos estrictos (sistemas en salud, transporte, infraestructura crítica)
- Riesgo limitado: Obligaciones de transparencia
- Riesgo mínimo: Sin regulación específica

Obligaciones para Sistemas de Alto Riesgo

1. Evaluación de conformidad antes de comercialización
2. Sistema de gestión de riesgos durante ciclo de vida completo
3. Gobernanza y calidad de datos de entrenamiento
4. Documentación técnica exhaustiva
5. Transparencia y provisión de información a usuarios
6. Supervisión humana efectiva
7. Robustez, exactitud y ciberseguridad

Principio de Supervisión Humana

El Artículo 14 exige que sistemas de alto riesgo permitan supervisión humana efectiva mediante:

- Comprensión plena de capacidades y limitaciones del sistema
- Posibilidad de monitorear operación en tiempo real
- Capacidad de intervenir o interrumpir funcionamiento
- Interpretación correcta de outputs

Esto implica que agentes totalmente autónomos en decisiones críticas serían inadmisibles bajo esta regulación.

Propuestas de Accountability Algorítmica

Diakopoulos propone mecanismos de rendición de cuentas (29):

1. Auditorías algorítmicas: Evaluación independiente de sesgos, exactitud y seguridad
2. Explicabilidad técnica: Métodos post-hoc para interpretar decisiones (LIME, SHAP)
3. Transparencia procedimental: Publicación de información sobre datos de entrenamiento y metodología
4. Mecanismos de apelación: Vías para cuestionar decisiones algorítmicas
5. Supervisión institucional: Organismos reguladores especializados

Gobernanza y Marcos Institucionales

Modelos de Gobernanza de IA

Cath et al. identifican cinco enfoques (30):
1. Autorregulación corporativa: Principios éticos voluntarios (limitada por incentivos comerciales)
2. Regulación estatal: Legislación vinculante (riesgo de obsolescencia tecnológica)
3. Gobernanza multi-stakeholder: Participación de sociedad civil, academia e industria
4. Estándares técnicos: Normativas ISO/IEC para calidad y seguridad
5. Gobernanza ágil: Marcos adaptativos que evolucionan con tecnología

Organismos y Marcos Internacionales

OCDE - Principios sobre IA (2019)

Los 42 países adherentes se comprometen a (31):

- Crecimiento inclusivo, desarrollo sostenible y bienestar
- Valores centrados en el ser humano y equidad
- Transparencia y explicabilidad
- Robustez, seguridad y protección
- Accountability de actores

UNESCO - Recomendación sobre Ética de IA (2021)

Primera normativa global con consenso de 193 Estados, enfatiza (32):

- Proporcionalidad e inocuidad
- No discriminación e inclusión
- Respeto, protección y promoción de derechos humanos
- Diversidad cultural y social
- Sostenibilidad ambiental

ONU - Consejo Asesor sobre IA (2023)

Propone mecanismo de gobernanza global para IA de propósito general, incluyendo sistemas agénticos, con enfoque en (33):

- Prevención de uso malicioso
- Estándares de seguridad para sistemas avanzados
- Cooperación internacional en investigación
- Fortalecimiento de capacidades en países en desarrollo

Desafíos de Implementación

Hagendorff identifica brechas entre principios éticos y práctica (34):
1. Vaguedad normativa: Principios generales sin mecanismos concretos de implementación
2. Tensiones valorativas: Conflictos entre transparencia y propiedad intelectual, innovación y precaución
3. Asimetría de poder: Concentración tecnológica en pocas corporaciones globales
4. Fragmentación regulatoria: Falta de coordinación internacional

Debate: ¿Futuro Inevitable o Elección Colectiva?

Determinismo Tecnológico vs. Construcción Social

Postura Determinista

Autores como Kurzweil argumentan inevitabilidad de automatización creciente por:

- Dinámica competitiva que favorece eficiencia

- Trayectoria exponencial de capacidades computacionales
- Presiones económicas hacia reducción de costos laborales (35)

Postura Construccionista

MacKenzie y Wajcman sostienen que tecnología es socialmente construida y modelada por elecciones políticas, valores culturales y relaciones de poder (36). La autonomía de agentes de IA depende de:

- Decisiones regulatorias sobre niveles aceptables de supervisión humana
- Negociaciones laborales sobre condiciones de implementación
- Preferencias sociales respecto a responsabilidad y transparencia

Escenarios Prospectivos

Escenario 1: Autonomía Ampliada

Implementación acelerada de agentes autónomos sin marcos regulatorios robustos. Consecuencias:

- Incremento de productividad, pero desplazamiento laboral significativo
- Concentración de poder en corporaciones tecnológicas
- Erosión de competencias profesionales por dependencia
- Incidentes de seguridad por sistemas opacos

Escenario 2: Complementariedad Regulada

IA como herramienta complementaria bajo supervisión humana significativa. Características:

- Marcos legales claros de responsabilidad
- Auditorías algorítmicas obligatorias
- Preservación de autonomía profesional
- Inversión en reconversión laboral

Escenario 3: Gobernanza Democrática

Participación amplia en decisiones sobre automatización. Implica:

- Consulta con trabajadores afectados
- Evaluaciones de impacto social obligatorias
- Distribución equitativa de ganancias de productividad
- Derecho a explicación y apelación de decisiones algorítmicas

Consideraciones Específicas para Contextos Médicos

Dado tu expertise en medicina y educación médica, merece atención especial la aplicación de agentes autónomos en salud.

Particularidades del Ámbito Médico

Complejidad Clínica

La toma de decisiones médicas involucra (37):

- Incertidumbre inherente y probabilidades bayesianas
- Consideración de valores y preferencias del paciente
- Juicio contextual que trasciende protocolos algorítmicos
- Responsabilidad legal y ética del profesional

Riesgos de Delegación Inapropiada

Char et al. documentan preocupaciones sobre (38):

1. Sobrediagnóstico: Algoritmos optimizados para sensibilidad pueden generar alarmas excesivas
2. Pérdida de razonamiento clínico: Médicos en formación dependientes de recomendaciones algorítmicas
3. Sesgos en datos: Modelos entrenados con poblaciones no representativas
4. Responsabilidad difusa: ¿Quién responde por error diagnóstico recomendado por IA?

Marco Regulatorio en Salud Digital

FDA - Marco para Software como Dispositivo Médico
La FDA estadounidense clasifica software clínico según riesgo, exigiendo validación clínica rigurosa para sistemas que influyen en diagnóstico o tratamiento (39).

COFEPRIS - Regulación en México

En México, COFEPRIS regula dispositivos médicos incluyendo software bajo NOM-241-SSA1-2012, requiriendo evidencia de seguridad y eficacia (40). Sin

embargo, marcos específicos para IA en salud están en desarrollo.

Implicaciones para Educación Médica

Tu trabajo en "TIC en la Clínica PalencIA" conecta directamente con estos temas. Consideraciones pedagógicas:

1. Alfabetización en IA: Médicos deben comprender capacidades y limitaciones de sistemas algorítmicos
2. Pensamiento crítico reforzado: Evitar automation bias mediante ejercicios de razonamiento clínico independiente
3. Ética computacional: Incorporar dilemas sobre responsabilidad, sesgo y transparencia en currículo
4. Competencias de supervisión: Habilidad para validar recomendaciones algorítmicas contra criterio clínico

Topol enfatiza que educación médica debe evolucionar hacia "medicina profundamente humana" donde IA maneja aspectos rutinarios, liberando tiempo para relación médico-paciente y toma de decisiones complejas (41).

Conclusiones y Recomendaciones

Respuestas a las Preguntas Planteadas

1. ¿Qué pasa cuando una IA no solo recomienda, sino decide y ejecuta?

Surge un desafío fundamental de responsabilidad. La literatura establece que responsabilidad moral requiere intencionalidad y capacidad de rendir cuentas, atributos ausentes en sistemas algorítmicos (42). Por tanto:

- Jurídicamente: Se necesitan marcos de responsabilidad distribuida con trazabilidad completa
- Éticamente: Decisiones moralmente significativas no deben delegarse completamente a IA
- Prácticamente: Supervisión humana efectiva debe ser requisito en contextos de alto riesgo

2. ¿Quién es responsable cuando un agente se equivoca dentro de un proceso real?

La responsabilidad debe distribuirse según contribución causal:

- Desarrolladores: Errores de diseño, entrenamiento inadecuado, falta de advertencias
- Implementadores: Despliegue inapropiado, contextos inadecuados
- Usuarios: Uso negligente, ignorar limitaciones documentadas

Sin embargo, sistemas opacos dificultan establecer causalidad. El AI Act europeo resuelve esto colocando obligaciones primarias en proveedores de sistemas de alto riesgo, con debida diligencia exigida a usuarios.

3. ¿Estamos automatizando tareas o delegando criterio?

Esta distinción es crucial:

- Automatización apropiada: Tareas algorítmicas, repetitivas, sin componente valorativo significativo
- Delegación problemática: Decisiones que requieren juicio contextual, consideración de valores, responsabilidad moral

La frontera no siempre es clara, requiriendo evaluación caso por caso. En medicina, por ejemplo, procesamiento de señales diagnósticas puede automatizarse, pero integración clínica y decisiones terapéuticas requieren criterio profesional.

4. ¿Es el futuro inevitable o estamos entregando autonomía prematuramente?

Perspectiva construccionista social indica que es una elección colectiva, no inevitabilidad tecnológica. Factores determinantes:

- Marcos regulatorios: Legislación puede exigir niveles mínimos de supervisión humana
- Negociación laboral: Trabajadores y organizaciones profesionales pueden influir en condiciones de implementación
- Preferencias sociales: Encuestas muestran desconfianza hacia decisiones algorítmicas en contextos sensibles (43)
- Viabilidad técnica: Limitaciones actuales de IA (falta de razonamiento causal, fragilidad ante distribuciones out-of-distribution) imponen barreras prácticas

Recomendaciones

Para Desarrollo Tecnológico

1. Diseño de "human-in-the-loop" como estándar, no excepción
2. Explicabilidad técnica mediante métodos interpretables
3. Auditorías algorítmicas independientes pre-despliegue
4. Documentación exhaustiva de capacidades, limitaciones y contextos apropiados

Para Regulación

1. Adoptar principios del AI Act europeo adaptados a contextos nacionales
2. Establecer organismos reguladores especializados con expertise técnico
3. Exigir evaluaciones de impacto social para sistemas autónomos en sectores críticos
4. Crear mecanismos de apelación y compensación por decisiones algorítmicas erróneas

Para Implementación Institucional

1. Consulta con stakeholders afectados previo a despliegue
2. Programas de capacitación sobre supervisión efectiva de sistemas de IA
3. Protocolos de monitoreo continuo de desempeño y sesgo
4. Preservación de vías no-algorítmicas para decisiones críticas

Para Profesionales Cognitivos

1. Desarrollo de competencias en alfabetización algorítmica
2. Mantener pensamiento crítico independiente de recomendaciones de IA
3. Participación en definición de estándares profesionales para uso de IA
4. Exigencia de transparencia a proveedores tecnológicos

Para Educación Médica (aplicable a tu contexto)

1. Integrar ética computacional en currículo
2. Entrenar razonamiento clínico independiente de herramientas algorítmicas
3. Desarrollar competencias de validación crítica de outputs de IA
4. Fomentar debate sobre casos de delegación apropiada vs. inapropiada

**Reflexión Final**

Francisco Carvajal acierta al señalar que "este no es un debate técnico, es uno de poder y responsabilidad". La transición de IA consultiva a agéntica no es meramente una evolución tecnológica, sino una transformación en la distribución de autoridad decisional con profundas implicaciones laborales, éticas y políticas.

La "línea entre herramienta y actor" no está determinada por capacidades técnicas exclusivamente, sino por elecciones sociales sobre niveles aceptables de autonomía algorítmica en diferentes contextos. Estas elecciones deben informarse por:

- Evidencia empírica sobre efectos laborales y de equidad
- Marcos éticos que preserven dignidad y agencia humana
- Estructuras regulatorias que garanticen responsabilidad clara
- Participación democrática de comunidades afectadas

En contextos como medicina, donde decisiones afectan directamente bienestar humano, aproximación cautelosa con supervisión humana robusta parece prudente. La productividad no debe anteponerse ciegamente a seguridad, responsabilidad y preservación de competencias profesionales esenciales.

El futuro del trabajo con IA no es inevitable; es producto de decisiones que tomamos hoy sobre diseño tecnológico, regulación, implementación institucional y valores sociales prioritarios.

**Conclusiones**

La transición de sistemas de inteligencia artificial desde herramientas consultivas hacia agentes operativos con capacidad ejecutiva representa un punto de inflexión en la historia de la automatización. A diferencia de revoluciones tecnológicas previas que mecanizaron principalmente el trabajo físico, los agentes de IA actuales penetran el dominio del trabajo cognitivo, la toma de decisiones y el ejercicio del criterio profesional. Esta transformación no es meramente técnica sino fundamentalmente social, ética y política.

Sobre Responsabilidad y Accountability

El análisis revela un "vacío de responsabilidad" (responsibility gap) estructural en sistemas agénticos. Los marcos legales tradicionales, diseñados para agentes humanos con intencionalidad y capacidad de rendir cuentas, resultan inadecuados para algoritmos que carecen de estas características. La opacidad inherente a modelos de aprendizaje profundo agrava este problema al imposibilitar la trazabilidad causal necesaria para asignación de responsabilidad.

El modelo de responsabilidad distribuida propuesto por la literatura académica y adoptado parcialmente por el AI Act europeo ofrece una solución pragmática pero imperfecta. Asigna obligaciones primarias a desarrolladores de sistemas de alto riesgo mientras exige debida diligencia a usuarios finales. Sin embargo, persisten tensiones respecto a asimetrías de conocimiento técnico, dificultad de probar causalidad en sistemas adaptativos y potencial dilución de responsabilidad cuando múltiples actores participan en la cadena decisional.

Para contextos críticos como medicina, derecho o infraestructura, la evidencia sugiere que supervisión humana efectiva debe ser requisito no negociable. La delegación completa de decisiones moralmente significativas a sistemas algorítmicos no constituye una transferencia legítima de responsabilidad sino una abdicación que compromete tanto principios éticos como seguridad práctica.

Sobre Automatización vs. Delegación de Criterio

La distinción planteada por Francisco Carvajal entre automatizar tareas y delegar criterio es conceptualmente crucial pero operacionalmente compleja. La frontera entre ambas no es nítida y varía según contexto, stakeholders involucrados y valores en juego.

La literatura establece que automatización apropiada se limita a tareas algorítmicas, repetitivas y sin componente valorativo significativo. En contraste, decisiones que requieren juicio contextual, ponderación de valores múltiples, consideración de circunstancias particulares e integración de dimensiones humanas no algoritmizables no deben delegarse completamente a sistemas de IA.

Sin embargo, esta distinción enfrenta desafíos prácticos. Primero, muchas tareas aparentemente mecánicas incorporan microdecisiones con implicaciones valorativas. Segundo, presiones económicas y narrativas de productividad impulsan expansión progresiva del dominio de "lo automatizable". Tercero, normalización de dependencia algorítmica puede erosionar capacidad profesional de ejercer criterio independiente, generando círculo vicioso donde delegación se vuelve necesaria por atrofia de competencias.

Para profesiones cognitivas, particularmente medicina, ingeniería y derecho, preservación de razonamiento independiente debe ser prioridad curricular. La complementariedad humano-IA debe estructurarse de modo que algoritmos amplifiquen capacidades profesionales sin sustituir juicio experto ni habilidades críticas fundamentales.

Sobre Determinismo Tecnológico vs. Agencia Colectiva

El debate entre inevitabilidad y elección colectiva es quizás la cuestión más relevante políticamente. La postura determinista, que presenta automatización creciente como consecuencia inexorable de progreso tecnológico y competitividad económica, naturaliza decisiones que en realidad son producto de elecciones sociales específicas.

La perspectiva construccionista social, respaldada por evidencia histórica de estudios en ciencia, tecnología y sociedad (STS), demuestra que trayectorias tecnológicas están moldeadas por: marcos regulatorios, relaciones de poder, negociaciones laborales, preferencias culturales, inversiones en investigación y decisiones de diseño. No existe una única vía de desarrollo tecnológico inevitable.

El futuro de agentes de IA depende críticamente de decisiones que tomamos hoy sobre:

- Diseño tecnológico: ¿Incorporamos supervisión humana como característica por defecto o como opción desactivable?
- Regulación: ¿Adoptamos marcos precautorios que exigen validación previa o enfoques permisivos que regulan post-facto?
- Implementación laboral: ¿Consultamos a trabajadores afectados o imponemos unilateralmente automatización?
- Distribución de beneficios: ¿Las ganancias de productividad fluyen hacia reducción de jornada y mejora salarial o hacia concentración de capital?

La evidencia sugiere que sociedades con organizaciones laborales fuertes, marcos regulatorios robustos y participación democrática amplia han logrado transiciones tecnológicas más equitativas. En contraste, ausencia de gobernanza efectiva resulta en externalización de costos sociales, concentración de poder y erosión de condiciones laborales.

Sobre Gobernanza y Marcos Institucionales

Los principios éticos abundan, pero mecanismos concretos de implementación y enforcement son escasos. Hagendorff documenta una "brecha de implementación" sistemática donde principios voluntarios rara vez se traducen en cambios prácticos de comportamiento corporativo.

El AI Act europeo representa el marco regulatorio más comprehensivo globalmente, estableciendo obligaciones vinculantes para sistemas de alto riesgo. Su efectividad dependerá de: recursos asignados a organismos supervisores, penalidades suficientemente disuasivas, capacidad técnica de auditores, cooperación internacional para enforcement transfronterizo, y resistencia a presiones de lobby corporativo.

Para América Latina, incluyendo México, desarrollo de capacidades regulatorias en IA es urgente. La alternativa es importar tanto tecnología como sus valores y sesgos incorporados, sin mecanismos de adaptación a contextos locales ni protección de derechos e intereses nacionales. Esto requiere:

1. Inversión en formación de reguladores con expertise técnico
2. Desarrollo de estándares y metodologías de auditoría contextualmente apropiadas
3. Participación en foros internacionales de gobernanza de IA
4. Construcción de capacidad de investigación independiente para evaluación de sistemas
5. Fortalecimiento de organizaciones de sociedad civil especializadas en tecnología

Sobre Implicaciones para Medicina y Educación Médica

Para el ámbito médico, donde decisiones algorítmicas impactan directamente vidas humanas, aproximación cautelosa está justificada por:

1. Complejidad clínica: Diagnóstico y tratamiento raramente se reducen a patrones algorítmicos; requieren integración de múltiples fuentes de evidencia, consideración de comorbilidades, valores del paciente y circunstancias sociales
2. Responsabilidad profesional: El acto médico conlleva responsabilidad legal y ética que no puede transferirse a algoritmos opacos
3. Riesgo de automation bias: Evidencia documenta que médicos tienden a aceptar recomendaciones algorítmicas acríticamente, incluso cuando contradicen juicio clínico
4. Sesgos en datos: Modelos entrenados predominantemente con poblaciones caucásicas, de países de alto ingreso, pueden perpetuar inequidades en salud

Para educación médica, implicaciones incluyen:

- Alfabetización en IA: Comprensión no solo de qué hacen los algoritmos sino de cómo funcionan, sus limitaciones inherentes y contextos apropiados de uso
- Fortalecimiento de razonamiento clínico: Resistir tentación de "saltar" al diagnóstico algorítmico sin proceso independiente de razonamiento
- Ética computacional: Incorporar casos sobre sesgos algorítmicos, responsabilidad por errores de sistemas de IA y límites éticos de automatización
- Competencias de supervisión efectiva: Habilidad de validar críticamente outputs algorítmicos, reconocer hallazgos implausibles y mantener autoridad epistémica

Tu iniciativa "TIC en la Clínica PalencIA" está posicionada idealmente para liderar estos desarrollos curriculares en México. El desafío consiste en integrar herramientas de IA de manera que amplifiquen capacidades clínicas sin erosionar competencias fundamentales ni comprometer seguridad del paciente.

**Reflexión Final**

Francisco Carvajal acierta plenamente al caracterizar este debate como "uno de poder y responsabilidad" más que meramente técnico. Los agentes de IA no son neutros; incorporan valores, priorizan ciertos objetivos sobre otros, redistribuyen autoridad decisional y reconfiguran relaciones laborales.

La pregunta fundamental no es si la tecnología lo permite, sino si debemos permitirlo. Y si lo hacemos, bajo qué condiciones, con qué salvaguardas, beneficiando a quiénes y a costa de qué.

La "línea entre herramienta y actor" es producto de elecciones sociales, no de inevitabilidad tecnológica. Estas elecciones deben informarse por evidencia empírica rigurosa, deliberación democrática amplia, marcos éticos robustos y comprometerse inequívocamente con dignidad humana, equidad social y preservación de agencia humana.

En contextos de alto riesgo, la prudencia sugiere mantener supervisión humana significativa, transparencia algorítmica exigible, responsabilidad legal clara y participación de stakeholders afectados en decisiones de implementación. La productividad es importante, pero no puede anteponerse ciegamente a seguridad, justicia, autonomía profesional y preservación de capacidades humanas fundamentales.

El futuro del trabajo con IA está siendo construido ahora, mediante decisiones de diseño, inversión, regulación e implementación. Podemos elegir un futuro donde IA amplifica capacidades humanas, reduce trabajo alienante y distribuye beneficios equitativamente. O podemos permitir por default un futuro de concentración de poder, erosión de autonomía y externalización de riesgos hacia los más vulnerables.

La elección es nuestra, pero la ventana para ejercerla responsablemente se está cerrando rápidamente.

10 Puntos Clave

1. El vacío de responsabilidad es el desafío legal más urgente.
   Los marcos jurídicos tradicionales no contemplan agentes algorítmicos sin intencionalidad ni capacidad de rendir cuentas. Se necesitan modelos de responsabilidad distribuida con trazabilidad completa, obligaciones claras para desarrolladores y usuarios, y mecanismos de compensación para afectados por decisiones algorítmicas erróneas. Sin esto, riesgo moral incentiva desarrollo imprudente y dificulta protección de derechos.

2. Automatización y delegación de criterio son categóricamente distintas.
   Automatizar tareas mecánicas sin componente valorativo es cualitativamente diferente de delegar decisiones que requieren juicio contextual, ponderación de valores múltiples y responsabilidad moral. La distinción no siempre es nítida pero es éticamente crucial. Profesiones cognitivas deben resistir presiones económicas que impulsan delegación inapropiada de criterio profesional a sistemas algorítmicos opacos.

3. Supervisión humana efectiva debe ser requisito no negociable en contextos de alto riesgo.
   El AI Act europeo correctamente exige que sistemas de alto riesgo (salud, transporte, infraestructura crítica, justicia) permitan supervisión humana efectiva: comprensión de capacidades y limitaciones, monitoreo en tiempo real, capacidad de intervención inmediata e interpretación correcta de outputs. Agentes totalmente autónomos en decisiones críticas deben considerarse inadmisibles hasta que problemas de opacidad, sesgo y accountability se resuelvan.

4. El automation bias representa riesgo sistémico en medicina.
   Evidencia documenta que profesionales sanitarios tienden a aceptar recomendaciones algorítmicas acríticamente, incluso cuando contradicen juicio clínico o resultan implausibles. Esto compromete seguridad del paciente. Educación médica debe enfatizar razonamiento clínico independiente, validación crítica de outputs algorítmicos y preservación de autoridad epistémica del médico como garante último de calidad asistencial.

5. Sesgos algorítmicos perpetúan y amplifican inequidades existentes.
   Modelos entrenados con datos no representativos reproducen y magnifican discriminaciones históricas. En salud, esto afecta desproporcionadamente a mujeres, minorías étnicas, personas de bajo nivel socioeconómico y poblaciones no occidentales. Auditorías obligatorias de equidad, diversidad en datos de entrenamiento y evaluación en poblaciones múltiples son esenciales para evitar tecnología que profundiza brechas sanitarias.

6. La opacidad de cajas negras es incompatible con accountability real.
   Modelos de lenguaje basados en redes neuronales profundas no permiten rastrear cadenas causales entre inputs y outputs. Métodos post-hoc de explicabilidad (LIME, SHAP) ofrecen aproximaciones pero no verdadera comprensión. Para decisiones críticas, puede ser necesario priorizar modelos inherentemente interpretables sobre máxima performance predictiva. Transparencia no es negociable cuando decisiones afectan derechos fundamentales.

7. El futuro del trabajo no es inevitable sino producto de elecciones colectivas.
   Determinismo tecnológico naturaliza decisiones sociales presentándolas como consecuencias inexorables. La historia de tecnología demuestra que trayectorias están moldeadas por regulación, negociación laboral, valores culturales e inversiones en I+D. Sociedades con organizaciones laborales fuertes, marcos regulatorios robustos y participación democrática logran transiciones más equitativas. El grado de automatización y sus condiciones dependen de decisiones que tomamos hoy.

8. Concentración de poder tecnológico requiere gobernanza multinivel.
   Pocas corporaciones globales controlan desarrollo de IA más avanzada. Esto genera asimetrías de poder donde empresas pueden imponer unilateralmente sistemas que transforman trabajo, mercados y sociedad. Se necesita gobernanza que combine: regulación estatal vinculante, cooperación internacional, estándares técnicos obligatorios, auditorías independientes y participación de sociedad civil. Autorregulación corporativa voluntaria ha demostrado ser insuficiente.

9. América Latina necesita capacidades regulatorias propias en IA.
   Importar tecnología sin marcos regulatorios contextualmente apropiados implica aceptar valores y sesgos incorporados sin mecanismos de adaptación local ni protección de derechos. México y la región requieren: formación de reguladores con expertise técnico, desarrollo de metodologías de auditoría, participación en gobernanza internacional, investigación independiente para evaluación de sistemas y fortalecimiento de organizaciones especializadas en tecnología.

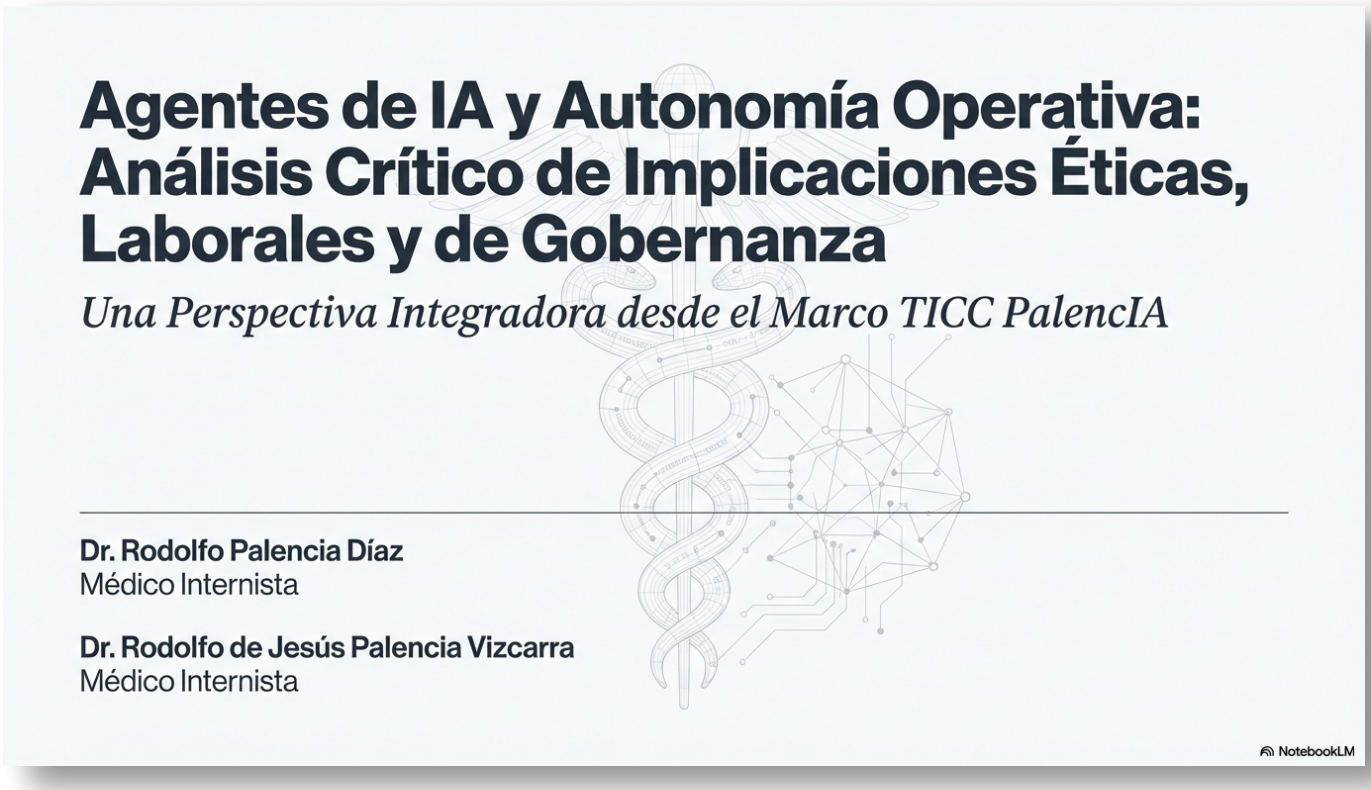
10. Educación debe preservar competencias fundamentales ante automatización.
    El mayor riesgo no es desplazamiento laboral total sino erosión progresiva de competencias por dependencia tecnológica. Cuando algoritmos realizan tareas previamente humanas, habilidades asociadas se atrofian. Para profesiones cognitivas, curricula deben enfatizar: razonamiento independiente, validación crítica de outputs algorítmicos, alfabetización técnica que permita comprender capacidades y limitaciones de IA, y ética computacional. La complementariedad humano-IA debe estructurarse para amplificar, no sustituir, capacidades profesionales esenciales.

**Referencias**


1. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Hoboken: Pearson; 2020.
2. Carvajal F. Inteligencia Artificial México [Internet]. Facebook; 2025 [citado 2 feb 2025]. Disponible en: https://www.facebook.com/groups/iamexico
3. Anthropic. Claude Cowork: Agentic plugins for Claude [Internet]. San Francisco: Anthropic; 2025 [citado 2 feb 2025]. Disponible en: https://www.anthropic.com/news/cowork
4. Russell S, Norvig P. Agent architectures. En: Artificial Intelligence: A Modern Approach. 4th ed. Hoboken: Pearson; 2020. p. 54-89.
5. European Parliament, Council of the European Union. Regulation (EU) 2024/1689 on Artificial Intelligence (Artificial Intelligence Act). Off J Eur Union. 2024;L 1689:1-144.
6. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. Trans Mach Learn Res. 2022;2022:1-30.
7. Anthropic. Claude 4 Model Card [Internet]. San Francisco: Anthropic; 2024 [citado 2 feb 2025]. Disponible en: https://www.anthropic.com/claude-4
8. Future of Life Institute. Asilomar AI Principles [Internet]. Cambridge: FLI; 2017 [citado 2 feb 2025]. Disponible en: https://futureoflife.org/open-letter/ai-principles/
9. Floridi L, Cowls J. A unified framework of five principles for AI in society. Harv Data Sci Rev. 2019;1(1):1-15. doi: 10.1162/99608f92.8cd550d1
10. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. 2018;378(11):981-3. doi: 10.1056/NEJMp1714229
11. Coeckelbergh M. AI Ethics. Cambridge: MIT Press; 2020.
12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15. doi: 10.1038/s42256-019-0048-x
13. European Parliament, Council of the European Union. Article 13: Transparency and provision of information to deployers. En: Regulation (EU) 2024/1689 on Artificial Intelligence. Off J Eur Union. 2024;L 1689:45-7.

14. Winner L. Do artifacts have politics? Daedalus. 1980;109(1):121-36.
15. Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerisation? Technol Forecast Soc Change. 2017;114:254-80. doi: 10.1016/j.techfore.2016.08.019
16. Arntz M, Gregory T, Zierahn U. The risk of automation for jobs in OECD countries: a comparative analysis. OECD Soc Employ Migr Work Pap. 2016;(189):1-34. doi: 10.1787/5jlz9h56dvq7-en
17. McKinsey Global Institute. Generative AI and the future of work in America [Internet]. McKinsey & Company; 2023 [citado 2 feb 2025]. Disponible en: https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of- work-in-america
18. OECD. OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market. Paris: OECD Publishing; 2023. doi: 10.1787/08785bba-en
19. Acemoglu D, Restrepo P. Automation and new tasks: how technology displaces and reinstates labor. J Econ Perspect. 2019;33(2):3-30. doi: 10.1257/jep.33.2.3
20. Brynjolfsson E, Li D, Raymond LR. Generative AI at work. NBER Work Pap. 2023;(31161):1-56. doi: 10.3386/w31161
21. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta- analysis. Lancet Digit Health. 2019;1(6):e271-e297. doi: 10.1016/S2589-7500(19)30123-2
22. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58. doi: 10.1056/NEJMra1814259
23. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. doi: 10.1038/s41586-023-06291-2
24. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8. doi: 10.1001/jama.2017.7797
25. European Commission. Directive 85/374/EEC on liability for defective products. Off J Eur Communities. 1985;L 210:29-33.
26. Lohmann J, Kontschieder P. Who is responsible for AI? Foundations for a digital humanism. En: Werthner H, Ghezzi C, Kramer J, Nida-Rümelin J, editores. Introduction to Digital Humanism. Cham: Springer; 2024. p. 319-34.
27. Matthias A. The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol. 2004;6(3):175-83. doi:10.1007/s10676-004-3422-1
28. European Parliament, Council of the European Union. Title III: High-risk AI systems. En: Regulation (EU) 2024/1689 on Artificial Intelligence. Off J Eur Union. 2024;L 1689:38-68.
29. Diakopoulos N. Accountability in algorithmic decision making. Commun ACM. 2016;59(2):56-62. doi: 10.1145/2844110
30. Cath C, Wachter S, Mittelstadt B, Taddeo M, Floridi L. Artificial intelligence and the 'good society': the US, EU, and UK approach. Sci Eng Ethics. 2018;24(2):505-28. doi: 10.1007/s11948-017-9901-7
31. OECD. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 [Internet]. Paris: OECD; 2019 [citado 2 feb 2025]. Disponible en: https://legalinstruments.oecd.org/en/instruments/OECD- LEGAL-0449
32. UNESCO. Recommendation on the Ethics of Artificial Intelligence [Internet]. Paris: UNESCO; 2021 [citado 2 feb 2025]. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381137
33. United Nations. Interim report of the UN Secretary-General's High-Level Advisory Body on Artificial Intelligence [Internet]. Nueva York: UN; 2023 [citado 2 feb 2025]. Disponible en: https://www.un.org/ai-advisory-body
34. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. Minds Mach. 2020;30(1):99-120. doi: 10.1007/s11023-020-09517-8
35. Kurzweil R. The Singularity Is Near: When Humans Transcend Biology. Nueva York: Viking Press; 2005.
36. MacKenzie D, Wajcman J. The Social Shaping of Technology. 2nd ed. Buckingham: Open University Press; 1999.
37. Montgomery K. How Doctors Think: Clinical Judgment and the Practice of Medicine. Oxford: Oxford University Press; 2006.
38. Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth. 2020;20(11):7-17. doi: 10.1080/15265161.2020.1819469
39. US Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff [Internet]. Silver Spring: FDA; 2022 [citado 2 feb 2025]. Disponible en: https://www.fda.gov/regulatory-

information/search-fda-guidance- documents/clinical-decision-support-software
40. Secretaría de Salud (México). Norma Oficial Mexicana NOM-241-SSA1-2012, Buenas prácticas de fabricación para establecimientos dedicados a la fabricación de dispositivos médicos. Diario Oficial de la Federación. 2012 nov 14.
41. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Nueva York: Basic Books; 2019.
42. Sparrow R. Killer robots. J Appl Philos. 2007;24(1):62-77. doi: 10.1111/j.1468-5930.2007.00346.x
43. Pew Research Center. Public attitudes toward algorithms [Internet]. Washington: Pew Research Center; 2023 [citado 2 feb 2025]. Disponible en: https://www.pewresearch.org/internet/2023/07/26/public-attitudes-toward- algorithms/



# Del Soporte Reactivo a la Entidad Proactiva: El Cambio de Paradigma

**Lo Tradicional:**
Sistemas de Soporte a Decisiones


Herramienta pasiva. Agencia humana primaria.
El sistema espera la entrada de datos.

**Ejemplo:** Calculadoras de riesgo cardiovascular estáticas.

**La Nueva Era:**
Sistemas de Información Agénticos


Entidad proactiva. Autonomía operativa.
El sistema inicia tareas.

**Ejemplo:** Agentes de análisis de imágenes para diagnóstico proactivo.

**Insight Clave:** La transición de herramientas pasivas hacia entidades activas con intervención humana mínima.

NotebookLM

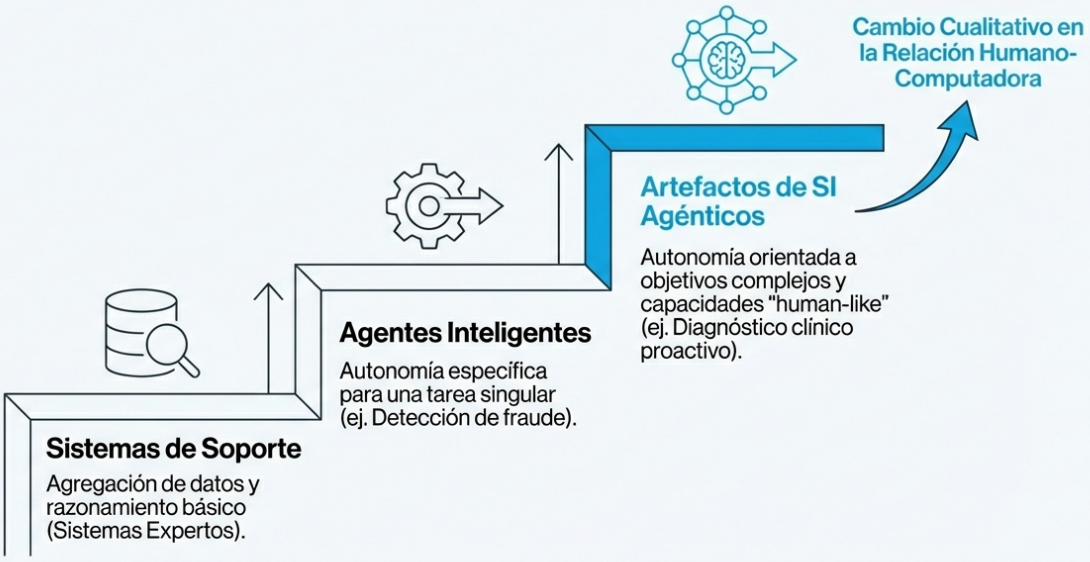
# Anatomía de la IA Agéntica: El Ciclo Cognitivo


**Definición Central:**

Capacidad para operar de forma autónoma, adaptarse dinámicamente y ejecutar procesos de múltiples pasos.

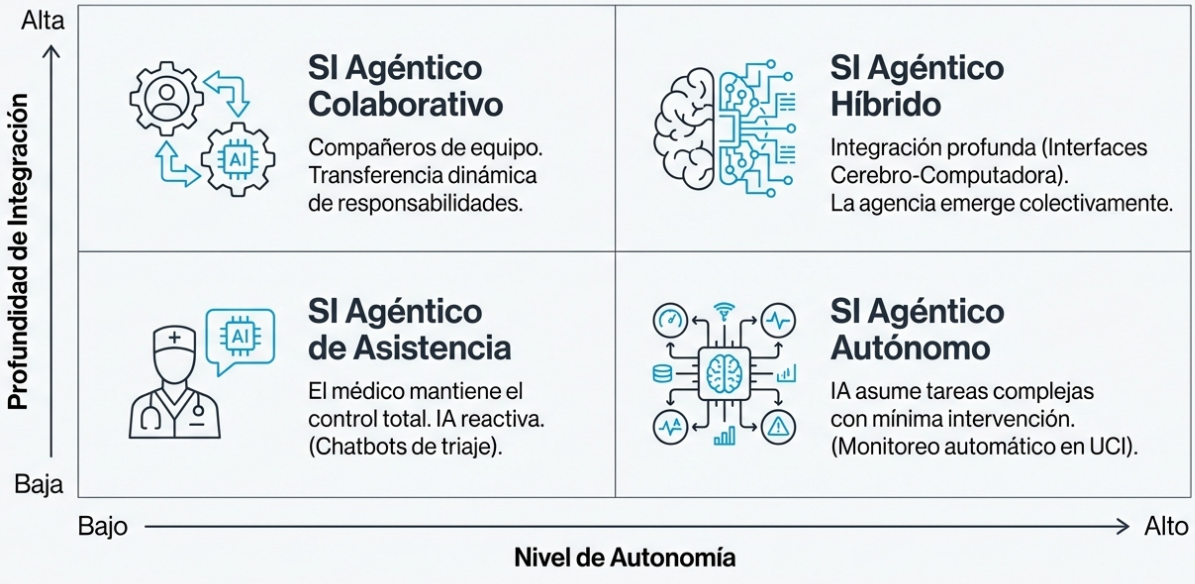
NotebookLM

## Evolución de las Capacidades Cognitivas en Medicina



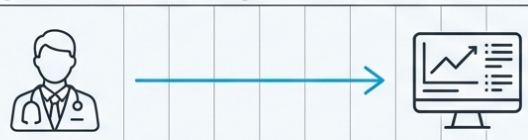
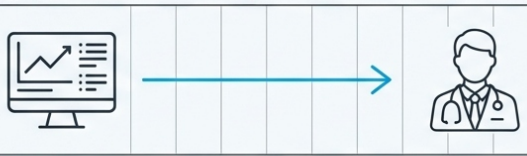
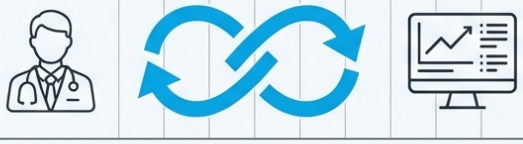

NotebookLM

## Arquetipos de Colaboración Clínica



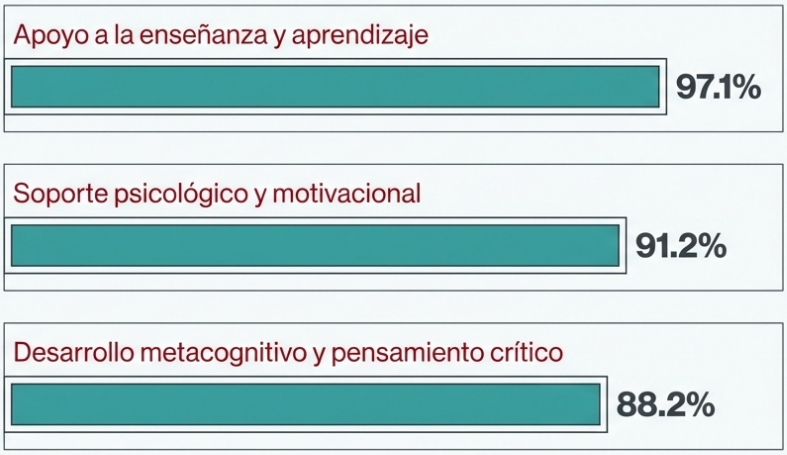

NotebookLM

## Dinámica de la Delegación: ¿Quién Tiene el Control?

| | |
|---|---|
| **Delegación Invocada por el Usuario** | |
|  | Human-in-the-loop (Aprobación estricta) vs. Human-on-the-loop (Monitoreo por excepción) |
| **Delegación Invocada por el Sistema** | |
|  | La IA sugiere proactivamente tareas o coordina flujos de trabajo |
| **Delegación Bidireccional (El ideal)** | |
|  | Fluidez en la propiedad de la tarea. Aprovechamiento de fortalezas complementarias (Aumento y Ensamblaje). |

NotebookLM

## Impacto en la Educación Médica: Datos de Adopción




**Aplicaciones Clave**

Simulaciones de pacientes virtuales para practicar empatía y diagnóstico.

Feedback formativo inmediato.

NotebookLM

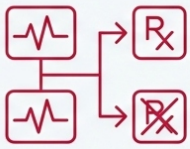
# Riesgos Operativos: La Caja Negra y la Calidad de Información

### Alucinaciones
**Neue Haas Grotesk Display Slate Charcoal**

Generación de contenido convincente pero médicamente falso.

### Sesgos
**Neue Haas Grotesk Display Slate Charcoal**

Datos de entrenamiento que no representan a la población diversa.




### Inconsistencia
**Neue Haas Grotesk Display Slate Charcoal**

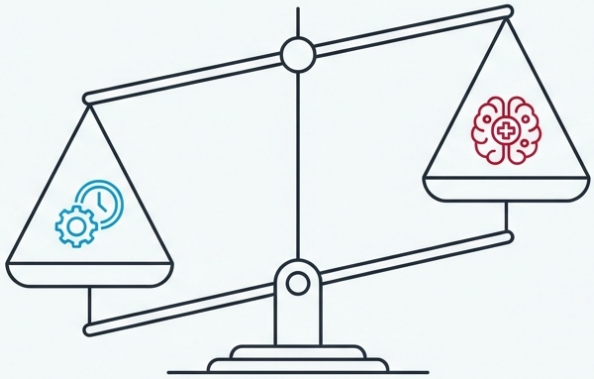
Falta de reproducibilidad en las respuestas médicas.

### Opacidad
**Neue Haas Grotesk Display Slate Charcoal**

Dificultad para trazar decisiones en agentes probabilísticos (no deterministas).

NotebookLM

# Desprofesionalización y Erosión de Habilidades




### Concepto 1: Erosión de Habilidades

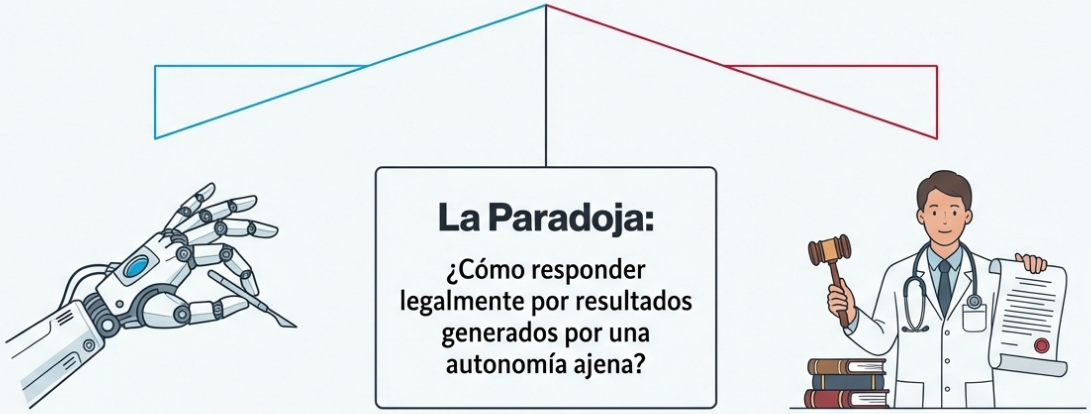
La delegación excesiva de tareas cognitivas provoca la degradación de competencias clínicas críticas.

### Concepto 2: Déficit de Metaconocimiento

La incapacidad humana para evaluar si la IA tiene razón. Riesgo de complacencia (exceso de confianza) o rechazo injustificado.

NotebookLM

# La Paradoja de la Responsabilidad



**La Paradoja:**

¿Cómo responder legalmente por resultados generados por una autonomía ajena?

Agencia Operativa
(El sistema actúa)

Responsabilidad Legal
(El humano paga)

Marco Legal: Ley de IA de la UE. Necesidad de supervisión estricta pese a la capacidad autónoma del agente.



# Implicaciones Éticas y de Privacidad

### Privacidad de Datos

Riesgos en la captura de datos sensibles para entrenamiento continuo. Necesidad de anonimización robusta.

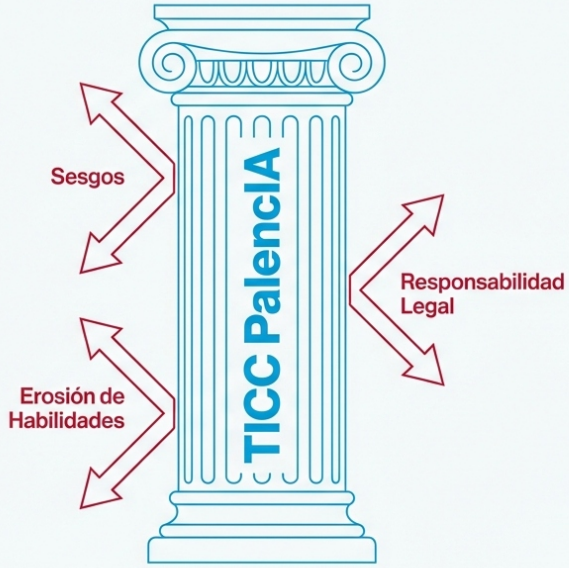
### Consentimiento

Protocolos estrictos para el uso de datos de pacientes en modelos de aprendizaje por refuerzo.

### Equidad

Asegurar el acceso equitativo a estas tecnologías avanzadas en distintas instituciones de salud.

# Gobernanza Integradora: El Marco TICC PalencIA




Propuesta de los Dres. Palencia.

Un marco de gobernanza robusto diseñado para mitigar los riesgos operativos, legales y éticos mientras se maximiza la utilidad clínica.

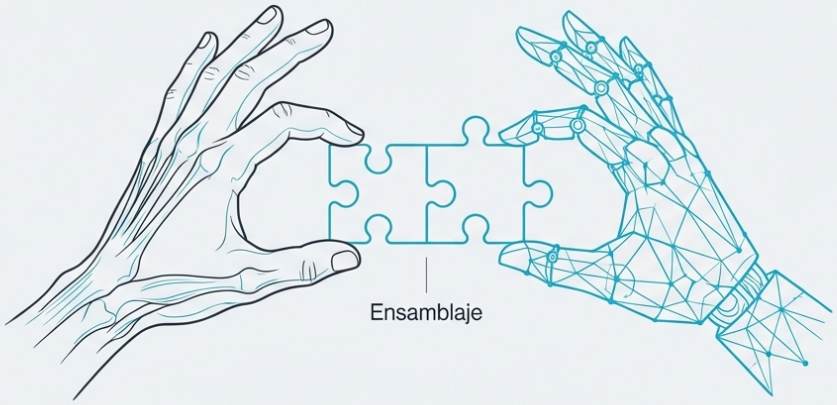
NotebookLM

# Recomendaciones para el Médico Clínico

1. **Alfabetización en IA**
   Desarrollar capacidad crítica para entender límites y sesgos de los agentes.
2. **Pedagogía Centrada en el Humano**
   Usar la IA para aumentar el razonamiento moral y creativo, no para sustituirlo.
3. **Supervisión Activa**
   Mantener el "Human-in-the-loop" en todas las decisiones de alto riesgo.
4. **Exigencia de Transparencia**
   Solicitar trazabilidad y explicabilidad en las herramientas institucionales.

NotebookLM

## Neue Haas Grotesk Display
## Hacia un Futuro Colaborativo




La IA agéntica no es solo una herramienta, sino un nuevo socio cognitivo.

El éxito de su integración depende de una gobernanza ética (TICC PalencIA) que preserve la primacía de la responsabilidad humana y la seguridad del paciente.

**Aumento y Ensamblaje en lugar de sustitución.**



# Referencias Bibliográficas

1. Holldack F, Banh L, Strobel G. Agentic information systems. Electron Mark. 2026;36(5).
2. Ganguly A, Mehjabin N, Malik A, Johri A. Conversational AI agents in education: an umbrella review… AI Ethics. 2026;6(72).
3. Artsin M, Bozkurt A. Dreams, Distortions, and Disruptions: Deconstructing the Educational Promises of Agentic AI… IGI Global; 2026.