THE SYNERGY HALL
SERIE DE INVESTIGACIÓN

UNA INICIATIVA DE LA FUNDACIÓN SHEYENE GERARDI
SEGURIDAD ESTRATÉGICA · CIENCIA · SOCIEDAD

N.º 3
Agosto de 2026
DOI

# El Marco de Confiabilidad C2S

## La Inteligencia Artificial Causal y el Apoyo a la Decisión Institucional en Dominios de Alta Consecuencia

*Por Sheyene Gerardi*

PUBLICADO POR LA FUNDACIÓN SHEYENE GERARDI
Investigación, Sistemas y Estudios Institucionales
Washington, D.C.

## SOBRE LA SERIE DE INVESTIGACIÓN DE LA SALA DE SINERGIA

La Serie de Investigación de la Sala de Sinergia es la serie de monografías académicas de la Fundación Sheyene Gerardi. Impulsa la investigación interdisciplinaria en diseño institucional, seguridad estratégica, inteligencia artificial, ciencia de sistemas, gobernanza y capacidad humana colectiva.

Cada volumen presenta un análisis original destinado a conectar el conocimiento científico, el desarrollo tecnológico y el diseño institucional.

La serie examina cómo los sistemas independientes pueden coordinarse intencionalmente para producir resultados mayores que la suma de sus contribuciones individuales.

Cada monografía se publica como un volumen numerado de forma independiente dentro de un programa de investigación continuo destinado a desarrollar modelos institucionales duraderos al servicio del florecimiento humano.

### DETALLES DE PUBLICACIÓN

**Título de la serie:** Serie de Investigación de la Sala de Sinergia
**Tipo de serie:** Monografías de Investigación Académica
**Editorial:** Fundación Sheyene Gerardi
**División de Investigación:** Investigación, Sistemas y Estudios Institucionales
**Lugar de publicación:** Washington, D.C., Estados Unidos
**Autora:** Sheyene Gerardi
**Periodicidad:** Irregular
**Idioma:** Español
**DOI:** *pendiente*
**Volumen Inaugural:** Monografía N.º 3, agosto de 2026
**Contacto:** contact@sheyene.org · www.sheyene.org

## DERECHOS DE AUTOR



## CITA SUGERIDA

Gerardi, S. (2026). *El Marco de Confiabilidad C2S: La Inteligencia Artificial Causal y el Apoyo a la Decisión Institucional en Dominios de Alta Consecuencia.* Synergy Hall Research Series, Monografía N.º 3. Washington, D.C.: Sheyene Gerardi Foundation.

# El Marco de Confiabilidad C2S

## La Inteligencia Artificial Causal y el Apoyo a la Decisión Institucional en Dominios de Alta Consecuencia

*Por Sheyene Gerardi*

Fundación Sheyene Gerardi · Washington, D.C. · agosto de 2026

---

*La inteligencia artificial se está convirtiendo en un componente esencial de la toma de decisiones institucional en los ámbitos de la seguridad, la salud, las finanzas y la administración pública. A medida que estos sistemas asumen roles consultivos cada vez más trascendentales, el desafío central ya no es únicamente la capacidad computacional, sino la confianza institucional.*

## Resumen Ejecutivo

La inteligencia artificial se está convirtiendo en un componente esencial de la toma de decisiones institucional en los ámbitos de la seguridad, la salud, las finanzas y la administración pública. A medida que estos sistemas asumen roles consultivos cada vez más trascendentales, el desafío central ya no es únicamente la capacidad computacional, sino la confianza institucional. Las decisiones que afectan vidas humanas, la seguridad nacional y la infraestructura estratégica requieren sistemas analíticos que no solo sean precisos, sino también transparentes, explicables y científicamente responsables.

Esta monografía presenta el **Marco de Confiabilidad C2S**, un modelo conceptual que propone que la inteligencia artificial causal puede fortalecer la confiabilidad del personal en dominios de alta consecuencia al apoyar —y no reemplazar— el juicio institucional humano. A partir de los principios establecidos en *SynsID* y *El Plan Maestro de Supervivencia*, el marco sostiene que la inteligencia artificial debe funcionar como un sistema de integración de evidencia capaz de organizar información conductual, del desarrollo, fisiológica y operativa compleja en modelos transparentes de apoyo a la decisión.

Se presta especial atención al surgimiento de modelos fundacionales capaces de interpretar información biológica, entre ellos la familia **Cell2Sentence (C2S-Scale)**, desarrollada por la Universidad de Yale, Google Research y Google DeepMind. Estas tecnologías demuestran la creciente capacidad de los grandes modelos de lenguaje para analizar sistemas biológicos complejos. Su importancia dentro de esta monografía, sin embargo, no radica en la predicción conductual directa, sino en ilustrar cómo los futuros sistemas de IA multimodal pueden contribuir a marcos de evaluación institucional más amplios.

El Marco de Confiabilidad C2S rechaza, por lo tanto, tanto el determinismo biológico como la toma de decisiones plenamente autónoma. En cambio, propone un modelo centrado en el ser humano, en el cual la IA causal explicable asiste a profesionales calificados mediante la integración de múltiples fuentes independientes de evidencia objetiva, preservando la rendición de cuentas, la transparencia, las salvaguardas éticas y el debido proceso.

Este marco se presenta como una contribución conceptual destinada a estimular el debate interdisciplinario sobre la futura relación entre la inteligencia artificial, la gobernanza institucional y la confiabilidad humana en entornos donde las decisiones individuales pueden acarrear consecuencias irreversibles.

# Parte I. El Fundamento Científico

La inteligencia artificial no puede mejorar el juicio institucional a menos que la ciencia en la que se apoya sea, en sí misma, sólida. Antes de que la IA pueda ayudar a evaluar la confiabilidad humana, debe operar sobre una comprensión precisa de cómo emerge el comportamiento humano.

La genética conductual, la neurociencia y la psicología contemporáneas demuestran que el carácter no es aleatorio ni está genéticamente predeterminado. Por el contrario, el comportamiento humano se desarrolla mediante la interacción continua entre la biología, el desarrollo, el entorno, la educación, la experiencia y la agencia individual. La inteligencia artificial debe, por lo tanto, integrar estas dimensiones en lugar de reducirlas a variables aisladas.

Las secciones siguientes establecen el fundamento científico sobre el cual se construye el Marco de Confiabilidad C2S.

## 1.1 El Carácter Humano como Sistema Adaptativo Complejo

La ciencia conductual moderna comprende cada vez más a la personalidad como un sistema adaptativo complejo, y no como el producto de factores biológicos aislados. Los estudios de gemelos demuestran de manera consistente que el carácter es moderadamente heredable entre poblaciones, mientras que la genética molecular revela que esta heredabilidad emerge de redes biológicas altamente distribuidas que involucran a cientos de variantes genéticas interactuantes, y no a genes individuales.[1]

Estos descubrimientos resuelven un importante malentendido científico. El comportamiento humano no está controlado por "genes de la agresión", "genes de la honestidad" ni por ningún otro determinante conductual aislado. Por el contrario, las predisposiciones biológicas contribuyen de manera probabilística al desarrollo de sistemas neuronales cuya expresión permanece continuamente influida por la educación, la cultura, las relaciones, la salud y la experiencia vivida.

En consecuencia, el objetivo de la evaluación institucional no puede ser la predicción del comportamiento a partir de la biología heredada. Más bien, la ciencia biológica contribuye a

comprender los orígenes del comportamiento humano, mientras que el desempeño observable sigue siendo el indicador más confiable de la confiabilidad operativa.

Esta distinción constituye el puente conceptual entre la ciencia conductual contemporánea y los principios institucionales desarrollados a lo largo de las monografías anteriores.

## 1.2 De la Información Biológica al Conocimiento Institucional

El conocimiento científico adquiere valor institucional únicamente cuando mejora la toma de decisiones sin exceder los límites de la evidencia disponible.

Los avances recientes en biología computacional —entre ellos los análisis a gran escala de redes poligénicas y los modelos fundacionales capaces de interpretar datos biológicos complejos— han ampliado sustancialmente la capacidad de la humanidad para comprender los sistemas vivos. Estas tecnologías demuestran una capacidad analítica extraordinaria, al tiempo que ilustran un principio igualmente importante: un mayor poder computacional no elimina la incertidumbre científica.

Dentro del Marco de Confiabilidad C2S, la información biológica funciona, por lo tanto, como un componente dentro de un ecosistema de evidencia más amplio. Aporta contexto y no conclusiones, nutre el juicio sin reemplazarlo, y fortalece la comprensión institucional sin convertirse en una base independiente para las decisiones de certificación.

Este principio preserva tanto la integridad científica como la legitimidad ética. Las instituciones se benefician de los avances de la ciencia biológica evitando, al mismo tiempo, los errores del determinismo biológico, el exceso de confianza tecnológica y los enfoques reduccionistas del comportamiento humano.

## 1.3 Por Qué la Inteligencia Artificial Requiere Causalidad

La creciente adopción de la inteligencia artificial en la toma de decisiones institucional ha expuesto una limitación fundamental de los modelos predictivos convencionales. La predicción estadística por sí sola no puede sustentar adecuadamente decisiones cuyas consecuencias involucran la seguridad nacional, los derechos humanos o un daño público irreversible.

La mayoría de los sistemas contemporáneos de aprendizaje automático identifican patrones a través de la correlación. Aunque a menudo son altamente precisos, estos modelos rara vez explican por qué se generó una predicción o qué mecanismos subyacentes produjeron el resultado observado. En los dominios de alta consecuencia, la predicción sin explicación resulta insuficiente para la rendición de cuentas institucional.

El Marco de Confiabilidad C2S distingue, por lo tanto, entre la inteligencia predictiva y la inteligencia causal. Los sistemas predictivos estiman la probabilidad de que ocurra un evento. Los sistemas causales buscan explicar las relaciones que generan esos resultados, evaluar escenarios alternativos y respaldar la intervención basada en evidencia.

Esta distinción cambia fundamentalmente el papel institucional de la inteligencia artificial. En lugar de funcionar como un evaluador automatizado de seres humanos, la IA se convierte en un instrumento analítico que ayuda a los profesionales a comprender interacciones complejas entre la evidencia biológica, conductual, del desarrollo, ambiental y operativa.

El resto de esta monografía explora cómo la IA causal puede fortalecer el juicio institucional preservando, al mismo tiempo, la transparencia, la rendición de cuentas y una supervisión humana significativa.

# Parte II. La Inteligencia Artificial Causal

La inteligencia artificial suele evaluarse según su precisión predictiva. Sin embargo, para la toma de decisiones institucional, la precisión por sí sola resulta insuficiente. Las decisiones que afectan la seguridad pública, la seguridad nacional o los derechos individuales requieren sistemas capaces de explicar sus conclusiones, identificar la incertidumbre y respaldar un juicio humano responsable.

La inteligencia artificial causal representa una evolución importante más allá de la analítica predictiva convencional. En lugar de limitarse a identificar asociaciones estadísticas, los modelos causales buscan representar los mecanismos a través de los cuales emergen los resultados, permitiendo a las instituciones evaluar la evidencia con mayor transparencia y explorar posibles intervenciones antes de que ocurran consecuencias irreversibles.

Las secciones siguientes examinan cómo el razonamiento causal puede contribuir a la confiabilidad institucional, permaneciendo al mismo tiempo sujeto a límites científicos, éticos y legales.

## 2.1 Más Allá de la Predicción

La inteligencia artificial predictiva ha transformado numerosos campos al identificar patrones dentro de grandes conjuntos de datos. La detección de fraude financiero, las imágenes médicas, la traducción de idiomas y muchas otras aplicaciones demuestran las notables capacidades del aprendizaje automático moderno. Sin embargo, la predicción por sí sola aporta solo una parte de la información necesaria para una toma de decisiones institucional responsable.

Los entornos de alta consecuencia exigen un nivel adicional de comprensión. Las instituciones deben saber no solo que se ha detectado un patrón determinado, sino también cómo se llegó a esa conclusión, qué variables contribuyeron de manera más significativa, qué incertidumbres persisten y si intervenciones alternativas podrían modificar el resultado proyectado.

La inteligencia artificial causal responde a estos requisitos al modelar explícitamente las relaciones entre variables, en lugar de basarse únicamente en la asociación estadística. A través de grafos causales, modelos causales estructurales y razonamiento contrafáctico, la IA puede ayudar a quienes toman las decisiones a distinguir las causas probables de las correlaciones coincidentales.

Dentro del Marco de Confiabilidad C2S, esta capacidad transforma a la inteligencia artificial de una tecnología predictiva en un sistema institucional de apoyo a la decisión, cuyo propósito principal no es el juicio autónomo, sino una mejor comprensión humana.

## 2.2 La Explicabilidad como Requisito Institucional

Las decisiones institucionales deben ser comprensibles tanto para los individuos afectados por ellas como para quienes son responsables de revisar, auditar y gobernar el proceso de toma de decisiones. La explicabilidad, por lo tanto, no es simplemente una característica técnica deseable de la inteligencia artificial; es un requisito previo para la legitimidad institucional.

Los modelos de caja negra pueden alcanzar un desempeño predictivo impresionante, aportando al mismo tiempo escasa comprensión sobre cómo se generan sus conclusiones. Esta opacidad puede resultar aceptable en aplicaciones donde las predicciones incorrectas conllevan consecuencias limitadas. Sin embargo, se vuelve inaceptable cuando la IA contribuye a decisiones que involucran la seguridad nacional, la certificación de personal u otras responsabilidades con resultados potencialmente irreversibles.

El Marco de Confiabilidad C2S exige, por lo tanto, que las evaluaciones asistidas por IA permanezcan transparentes, trazables y científicamente interpretables. Los revisores humanos deben ser capaces de comprender la evidencia integrada por el sistema, identificar los factores principales que influyen en cada recomendación, y cuestionar las conclusiones cuando sea necesario.

La inteligencia artificial debe, por lo tanto, funcionar como un colaborador analítico y no como una autoridad incuestionable. La explicabilidad transforma la capacidad computacional en confianza institucional, al garantizar que toda recomendación permanezca abierta al escrutinio científico, la evaluación ética y la rendición de cuentas humana.

## 2.3 Cell2Sentence como Estudio de Caso

Los avances recientes en los modelos fundacionales biológicos ilustran la rapidez con la que la inteligencia artificial está expandiendo su capacidad para interpretar información científica altamente compleja. Entre los ejemplos más significativos se encuentra la familia **Cell2Sentence (C2S-Scale)** de modelos, desarrollada mediante la colaboración entre la Universidad de Yale, Google Research y Google DeepMind.

Estos modelos transforman los datos de expresión génica de célula única en representaciones que los grandes modelos de lenguaje pueden analizar, lo que permite integrar información biológica compleja para tareas como la identificación de tipos celulares, la predicción de perturbaciones y la generación de hipótesis biomédicas. Su éxito demostrado dentro de la biología celular representa un hito importante en la aplicación de modelos fundacionales a las ciencias de la vida.

Dentro del Marco de Confiabilidad C2S, sin embargo, Cell2Sentence no se presenta como un sistema de evaluación conductual. Más bien, sirve como un ejemplo ilustrativo de una transición tecnológica más amplia: el surgimiento de una inteligencia artificial capaz de integrar información biológica a una escala sin precedentes.

Esta distinción es esencial. El marco no propone que los modelos de lenguaje biológico existentes puedan determinar la confiabilidad humana. Más bien, sostiene que los avances representados por tecnologías como Cell2Sentence ofrecen una perspectiva sobre cómo los futuros sistemas de IA multimodal pueden contribuir al apoyo a la decisión institucional cuando se integran responsablemente con la ciencia conductual, la neurociencia, la psicología y la evidencia longitudinal objetiva.

En consecuencia, Cell2Sentence funciona dentro de esta monografía como un estudio de caso científico, y no como el fundamento del propio marco de certificación propuesto.

### 2.4 Los Límites de la Inteligencia Artificial

La inteligencia artificial continúa avanzando a un ritmo extraordinario. No obstante, todo sistema de IA sigue estando limitado por la calidad de sus datos subyacentes, los supuestos incorporados en sus modelos y la validez científica de la evidencia de la cual dependen sus conclusiones.

Estas limitaciones adquieren especial importancia en la evaluación del comportamiento humano. La confiabilidad conductual no puede reducirse a mediciones biológicas aisladas, registros históricos o predicciones computacionales. La conducta humana sigue siendo dinámica, dependiente del contexto y continuamente influida por nuevas experiencias, entornos cambiantes, la educación, la salud y el desarrollo personal.

Por esta razón, el Marco de Confiabilidad C2S rechaza explícitamente la certificación autónoma por IA. La inteligencia artificial puede organizar evidencia, identificar relaciones, estimar la incertidumbre y asistir en la evaluación profesional, pero la responsabilidad institucional debe seguir recayendo en tomadores de decisiones humanos responsables, que operen bajo estándares legales y éticos transparentes.

El objetivo, por lo tanto, no es reemplazar el juicio humano por máquinas, sino mejorar la calidad, la consistencia y el fundamento científico del juicio institucional mediante una inteligencia artificial gobernada de manera responsable.

## Parte III. El Marco de Confiabilidad C2S

La inteligencia artificial alcanza su mayor valor institucional no cuando reemplaza el juicio humano, sino cuando lo fortalece.

El Marco de Confiabilidad C2S se propone como un modelo conceptual mediante el cual la inteligencia artificial causal asiste a la toma de decisiones institucional, integrando diversas fuentes de evidencia científicamente relevante en evaluaciones transparentes, explicables y

responsables. Su propósito no es automatizar la certificación, sino mejorar la calidad y la consistencia del juicio institucional humano.

El marco se construye sobre dos principios establecidos en las monografías anteriores. Primero, la certificación confiable debe evaluar al individuo expresado y no la predisposición biológica heredada. Segundo, la confiabilidad institucional aumenta a medida que las fuentes independientes de evidencia se integran deliberadamente, en lugar de evaluarse de forma aislada.

Las secciones siguientes describen cómo estos principios pueden operacionalizarse a través de un modelo centrado en el ser humano de apoyo a la decisión institucional asistido por IA.

## 3.1 Principios del Juicio Institucional Asistido por IA

El Marco de Confiabilidad C2S se organiza en torno a un principio simple: la inteligencia artificial debe ayudar a las instituciones a integrar evidencia, no reemplazar la responsabilidad institucional.

Toda recomendación generada por un sistema de IA debe seguir siendo comprensible, revisable de forma independiente y estar respaldada por evidencia objetiva. Quienes toman las decisiones humanas conservan la responsabilidad plena de evaluar las recomendaciones, considerar información contextual no disponible para los modelos computacionales, y garantizar que cada decisión siga siendo ética y legalmente defendible.

Este principio distingue el apoyo a la decisión institucional de la toma de decisiones automatizada. La inteligencia artificial aporta capacidad computacional, consistencia y profundidad analítica, mientras que las instituciones humanas aportan juicio, rendición de cuentas, proporcionalidad y responsabilidad moral.

La confianza institucional surge, por lo tanto, no de la inteligencia artificial por sí sola, sino de la integración deliberada entre la capacidad computacional y una gobernanza humana responsable.

## 3.2 El Modelo de Integración de Evidencia

En lugar de evaluar indicadores aislados, el Marco de Confiabilidad C2S organiza múltiples fuentes independientes de información en una arquitectura de evidencia unificada.

Estas fuentes pueden incluir la historia de desarrollo verificada, la evaluación conductual, el desempeño profesional, la evaluación psicológica, el monitoreo fisiológico, los registros operativos y —cuando resulte científicamente pertinente— información biológica validada. Ninguna categoría individual funciona como un determinante independiente de las decisiones institucionales.

La inteligencia artificial contribuye al identificar relaciones entre estas formas heterogéneas de evidencia, señalar inconsistencias, estimar la incertidumbre y presentar resúmenes analíticos transparentes para la revisión humana.

El objetivo no es la predicción por sí misma. El objetivo es mejorar la comprensión institucional, garantizando que la evidencia compleja se evalúe de manera consistente, integral y proporcional a las responsabilidades confiadas a cada individuo.

### 3.3 La Supervisión Humana y la Responsabilidad Institucional

La integración de la inteligencia artificial en la toma de decisiones institucional no disminuye la responsabilidad humana: la aumenta. A medida que los sistemas computacionales adquieren mayor capacidad para organizar información compleja, las instituciones asumen una mayor obligación de garantizar que cada recomendación siga siendo científicamente justificada, éticamente defendible y legalmente responsable.

Dentro del Marco de Confiabilidad C2S, la inteligencia artificial nunca funciona como la instancia decisoria final. Los profesionales humanos siguen siendo responsables de interpretar la evidencia, considerar factores contextuales que exceden el análisis computacional, y ejercer un juicio coherente con los valores institucionales y los estándares legales.

Este principio preserva una distinción fundamental entre inteligencia y autoridad. La inteligencia artificial puede aportar capacidad analítica, pero la autoridad institucional debe seguir siendo inseparable de la rendición de cuentas humana. Las decisiones que afectan carreras, la seguridad pública o la seguridad nacional requieren, en última instancia, una responsabilidad moral que no puede delegarse a sistemas computacionales.

En consecuencia, el marco concibe a la IA como un colaborador permanente en el razonamiento institucional, y no como un actor institucional autónomo.

## Parte IV. Gobernanza Ética e Institucional

La inteligencia artificial no puede fortalecer la confianza pública a menos que las instituciones que gobiernan su uso sean, ellas mismas, dignas de confianza. La capacidad científica por sí sola es insuficiente. La toma de decisiones institucional legítima requiere transparencia, rendición de cuentas, proporcionalidad, debido proceso y respeto por la dignidad humana.

El Marco de Confiabilidad C2S sitúa, por lo tanto, a la gobernanza en el centro del apoyo a la decisión asistido por IA. Las salvaguardas éticas no son restricciones secundarias impuestas sobre los sistemas tecnológicos; constituyen una parte del propio marco.

Los siguientes principios establecen las condiciones institucionales necesarias para la integración responsable de la inteligencia artificial en la toma de decisiones de alta consecuencia.

### 4.1 De la Disuasión a la Certificación

Las decisiones institucionales deben seguir siendo comprensibles para quienes son responsables de tomarlas, revisarlas y vivir con sus consecuencias. La inteligencia artificial tiene, por lo tanto, la obligación de producir explicaciones y no únicamente conclusiones.

La explicabilidad cumple varios propósitos institucionales de manera simultánea. Permite la validación científica, respalda la revisión legal, facilita la auditoría independiente, refuerza la confianza pública y permite que los individuos afectados por las decisiones institucionales comprendan la evidencia que las respalda.

La transparencia también protege a las propias instituciones. Los sistemas cuyo razonamiento no puede examinarse no pueden corregirse eficazmente cuando ocurren errores. La IA explicable, por lo tanto, mejora no solo la equidad, sino también la resiliencia institucional, al permitir un perfeccionamiento continuo a medida que evoluciona el conocimiento científico.

Dentro del Marco de Confiabilidad C2S, la transparencia no se trata como una característica técnica deseable. Es un principio institucional rector.

## 4.2 La Ética Más Allá del Cumplimiento Normativo

La gobernanza ética exige más que el cumplimiento de la normativa vigente. Las leyes establecen estándares mínimos de conducta aceptable, mientras que la ética institucional busca garantizar que la capacidad científica permanezca alineada con la dignidad humana y el interés público más amplio.

El Marco de Confiabilidad C2S rechaza, por lo tanto, los enfoques basados en la vigilancia, el determinismo biológico, la autoridad algorítmica opaca o la clasificación irreversible. En cambio, enfatiza la evaluación proporcional, la equidad procesal, la revisión continua y oportunidades reales de supervisión humana y apelación.

La ética se entiende, en consecuencia, como un componente habilitador de la confiabilidad institucional, y no como una limitación externa a la innovación tecnológica. Una gobernanza responsable fortalece la confianza pública, preservando al mismo tiempo la legitimidad necesaria para la eficacia institucional a largo plazo.

## 4.3 La Gobernanza Internacional y los Estándares Compartidos

Muchas de las instituciones responsables de gestionar riesgos de alta consecuencia operan a través de fronteras nacionales. La seguridad nuclear, la aviación civil, la bioseguridad, la inteligencia artificial, la ciberseguridad y la infraestructura crítica dependen cada vez más de la cooperación entre gobiernos, instituciones científicas y organizaciones internacionales.

El Marco de Confiabilidad C2S se presenta, por lo tanto, como un fundamento conceptual para el diálogo internacional y no como una propuesta de autoridad global centralizada. Las naciones individuales conservan la responsabilidad plena de implementar políticas coherentes con sus tradiciones constitucionales, sistemas legales y prioridades institucionales.

No obstante, la cooperación internacional ofrece ventajas sustanciales. Los estándares científicos compartidos, las metodologías de evaluación interoperables, la auditoría independiente, la investigación colaborativa y los mecanismos de gobernanza transparente

pueden fortalecer la confiabilidad institucional preservando, al mismo tiempo, la soberanía nacional.

El objetivo no es la uniformidad institucional, sino la compatibilidad institucional. La cooperación confiable surge cuando instituciones independientes operan conforme a principios compartidos de integridad científica, transparencia, rendición de cuentas y respeto por la dignidad humana.

## Conclusión

La inteligencia artificial representa una de las tecnologías institucionales más significativas del siglo XXI. Su mayor contribución, sin embargo, no surgirá de reemplazar el juicio humano, sino de mejorar la calidad de la toma de decisiones institucional mediante la integración responsable del conocimiento científico.

El Marco de Confiabilidad C2S propone que la inteligencia artificial causal funcione como un sistema institucional de apoyo a la decisión, y no como un agente decisorio autónomo. Al integrar la ciencia conductual, la neurociencia, la psicología, la evidencia operativa y —cuando resulte científicamente pertinente— información biológica validada en modelos analíticos transparentes, la IA puede fortalecer la consistencia, la rendición de cuentas y el fundamento científico de la evaluación del personal en dominios de alta consecuencia.

Al mismo tiempo, esta monografía enfatiza que la capacidad tecnológica debe permanecer subordinada a la responsabilidad institucional. La explicabilidad, la supervisión humana, el debido proceso, la proporcionalidad y la gobernanza ética no son salvaguardas opcionales añadidas después del desarrollo tecnológico; constituyen las condiciones bajo las cuales la inteligencia artificial se vuelve digna de la confianza institucional.

El marco aquí presentado se ofrece, por lo tanto, no como un programa operativo, sino como una contribución conceptual a la relación en evolución entre la inteligencia artificial y la gobernanza institucional. Su objetivo más amplio es fomentar la colaboración interdisciplinaria entre científicos, ingenieros, formuladores de políticas, juristas, especialistas en ética e instituciones públicas que buscan desarrollar modelos responsables de toma de decisiones asistida por IA.

En última instancia, el futuro de la inteligencia artificial estará determinado menos por la sofisticación de sus algoritmos que por la sabiduría de las instituciones que deciden cómo se utilizan esos algoritmos.

## Descargo de Responsabilidad

Esta monografía presenta un marco conceptual destinado a estimular el debate interdisciplinario sobre la integración responsable de la inteligencia artificial en la toma de decisiones institucional en dominios de alta consecuencia. No constituye política legal, gubernamental, regulatoria u operativa.

Las referencias a tecnologías de terceros —incluida la familia de modelos Cell2Sentence (C2S-Scale) desarrollada por la Universidad de Yale, Google Research y Google DeepMind— se limitan a sus capacidades científicas documentadas dentro de la biología celular. Cualquier análisis sobre su posible relevancia futura para la confiabilidad del personal o el apoyo a la decisión institucional representa el análisis conceptual de la autora y no debe interpretarse como un respaldo, afiliación o representación por parte de esas organizaciones.

El Marco de Confiabilidad C2S rechaza explícitamente el determinismo biológico y la toma de decisiones por IA plenamente autónoma. A lo largo de esta monografía, la inteligencia artificial se presenta exclusivamente como una tecnología de apoyo a la decisión, destinada a fortalecer el juicio institucional humano responsable, bajo las debidas salvaguardas científicas, éticas, legales y procesales.

Las opiniones expresadas son exclusivamente las de la autora y buscan fomentar el diálogo basado en evidencia sobre la inteligencia artificial, la gobernanza institucional y la confiabilidad humana.

## Obras Citadas


**1.** Zwir I, Arnedo J, Del-Val C, et al. Uncovering the complex genetics of human character. Molecular Psychiatry. 2020;25:2295–2312. Consultado el 2 de noviembre de 2025. https://www.nature.com/articles/s41380-018-0263-6

**2.** Zwir I, Arnedo J, Del-Val C, et al. Uncovering the complex genetics of human temperament. Molecular Psychiatry. 2020;25:2275–2294. Consultado el 2 de noviembre de 2025. https://www.nature.com/articles/s41380-018-0264-5

**3.** Caspi A, McClay J, Moffitt TE, et al. Role of genotype in the cycle of violence in maltreated children. Science. 2002;297:851–854. Consultado el 2 de noviembre de 2025. https://www.science.org/doi/10.1126/science.1072290

**4.** Byrd AL, Manuck SB. MAOA, childhood maltreatment, and antisocial behavior: meta-analysis of a gene-environment interaction. Biological Psychiatry. 2014;75(1):9–17. Consultado el 2 de noviembre de 2025. https://moffittcaspi.trinity.duke.edu/sites/moffittcaspi.trinity.duke.edu/files/file-attachments/MAOA_2013_ByrdManuck.pdf

**5.** Explaining heritable variance in human character (methodological commentary on the PGMRA model). bioRxiv preprint. 2018. Consultado el 2 de noviembre de 2025. https://www.biorxiv.org/content/10.1101/446518.full.pdf

**6.** Gene Expression Networks Regulated by Human Personality. PMC, National Institutes of Health. Consultado el 2 de noviembre de 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11408262/

**7.** Cell2Sentence-Scale Gemma 2 27B (C2S-Scale): model card. Yale University, Google Research, and Google DeepMind. Hugging Face. Consultado el 2 de noviembre de 2025. https://huggingface.co/vandijklab/C2S-Scale-Gemma-2-27B

**8.** Cell2Sentence: Teaching Large Language Models the Language of Biology. van Dijk Lab, Yale University. GitHub. Consultado el 2 de noviembre de 2025. https://github.com/vandijklab/cell2sentence

**9.** Google's Gemma AI Model Helps Discover a New Potential Cancer Therapy Pathway (Cell2Sentence-Scale 27B). Google. 2025. Consultado el 4 de junio de 2026. https://blog.google/innovation-and-ai/products/google-gemma-ai-cancer-therapy-discovery/

**10.** Bridging Biology and AI: Yale and Google's Collaborative Breakthrough in Single-Cell RNA Analysis. Yale School of Medicine. 2025. Consultado el 4 de junio de 2026. https://medicine.yale.edu/news-article/bridging-biology-and-ai-yale-and-googles-collaborative-breakthrough-in-single-cell-analysis/

**11.** The Explainability Challenge of Generative AI and LLMs. OCEG. Consultado el 2 de noviembre de 2025. https://www.oceg.org/the-explainability-challenge-of-generative-ai-and-llms/

**12.** LLMs for Explainable AI: A Comprehensive Survey. arXiv:2504.00125. Consultado el 2 de noviembre de 2025. https://arxiv.org/html/2504.00125v1

**13.** Harnessing AI in Multi-Modal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine. PMC, National Institutes of Health. Consultado el 2 de noviembre de 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11972123/

**14.** A Zero-Shot LLM-Based Framework for Descriptive Gene-Gene Interaction Network Generation. OpenReview. Consultado el 2 de noviembre de 2025. https://openreview.net/pdf?id=TKOklNCZNE

**15.** Personality Traits in Large Language Models. arXiv:2307.00184. Consultado el 2 de noviembre de 2025. https://arxiv.org/html/2307.00184v4

**16.** Risk Profiling and Modulation for LLMs. arXiv:2509.23058. Consultado el 2 de noviembre de 2025. https://arxiv.org/abs/2509.23058

**17.** Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. PMC, National Institutes of Health. Consultado el 4 de junio de 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC9828290/

**18.** Researchers Optimize Genetic Tests for Diverse Populations to Tackle Health Disparities. National Human Genome Research Institute. 2024. Consultado el 4 de junio de 2026. https://www.genome.gov/news/news-release/researchers-optimize-genetic-tests-for-diverse-populations-to-tackle-health-disparities

**19.** DoD Instruction 5210.42: DoD Nuclear Weapons Personnel Reliability Assurance. U.S. Department of Defense. Consultado el 4 de junio de 2026. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/521042p.pdf

**20.** Nuclear Security Series. International Atomic Energy Agency. Consultado el 4 de junio de 2026. https://www.iaea.org/resources/nuclear-security-series

**21.** Application of Machine Learning in Predicting Aggressive Behaviors from Hospitalized Patients with Schizophrenia. PMC, National Institutes of Health. Consultado el 4 de junio de 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC10067917/

**22.** Feng T, et al. Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems. arXiv:2601.22401. 2026. Consultado el 5 de junio de 2026. https://arxiv.org/abs/2601.22401