

METODOLOGÍA DE LA  
INVESTIGACIÓN CIENTÍFICA  
APLICADA A LA  
INTELIGENCIA ARTIFICIAL Y  
LA CIENCIA DE DATOS:  
**ENFOQUE GENERAL**



**Escrito por:**

Víctor Manuel Fernández López  
John Fernando Ruiz Alvarado  
Juan Sandoval Vilchez  
Oscar Adrian Zapillado Huanco  
Walter Jacinto Panocca Orellana  
Walter Hernan Patricio Peralta



## Metodología de la investigación científica aplicada a la inteligencia artificial y la ciencia de datos: Enfoque general

*Fernández López, Víctor Manuel; Ruiz Alvarado, John Fernando; Sandoval Vilchez, Juan; Zapillado Huanco, Oscar Adrian; Panocca Orellana, Walter Jacinto; Patricio Peralta, Walter Hernan*

© Fernández López, Víctor Manuel; Ruiz Alvarado, John Fernando; Sandoval Vilchez, Juan; Zapillado Huanco, Oscar Adrian; Panocca Orellana, Walter Jacinto; Patricio Peralta, Walter Hernan , 2025

Primera edición (1ra. ed.): Diciembre, 2025

Editado por:

**Editorial Mar Caribe ®**

[www.editorialmarcaribe.es](http://www.editorialmarcaribe.es)

Av. Gral. Flores 547, 70000 Col. del Sacramento, Departamento de Colonia, Uruguay.

Diseño de caratula e ilustraciones: *Isbelia Salazar Morote*

Libro electrónico disponible en:

<https://editorialmarcaribe.es/ark:/10951/isbn.9789915698540>

Formato: Electrónico

ISBN: 978-9915-698-54-0

ARK: [ark:/10951/isbn.9789915698540](https://nbn-resolving.org/urn:nbn:org:ark:10951-isbn.9789915698540)

[Editorial Mar Caribe \(OASPA\)](#): Como miembro de la Open Access Scholarly Publishing Association, apoyamos el acceso abierto de acuerdo con el código de conducta, transparencia y mejores prácticas de OASPA para la publicación de libros académicos y de investigación. Estamos comprometidos con los más altos estándares editoriales en ética y deontología, bajo la premisa de «Ciencia Abierta en América Latina y el Caribe»

# OASPA

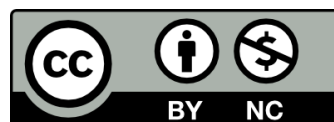
Editorial Mar Caribe, firmante N° 795 de 12.08.2024 de la [Declaración de Berlín](#)

"... Nos sentimos obligados a abordar los retos de Internet como medio funcional emergente para la distribución del conocimiento. Obviamente, estos avances pueden modificar significativamente la naturaleza de la publicación científica, así como el actual sistema de garantía de calidad...." (Max Planck Society, ed. 2003., pp. 152-153).



[CC BY-NC 4.0](#)

Los autores pueden autorizar al público en general a reutilizar sus obras únicamente con fines no lucrativos, los lectores pueden utilizar una obra para generar otra, siempre que se dé crédito a la investigación, y conceden al editor el derecho a publicar primero su ensayo bajo los términos de la licencia CC BY-NC 4.0.



Editorial Mar Caribe se adhiere a la "Recomendación relativa a la preservación del patrimonio documental, comprendido el patrimonio digital, y el acceso al mismo" de la UNESCO y a la Norma Internacional de referencia para un sistema abierto de información archivística ([OAIS-ISO 14721](#)). Este libro está preservado digitalmente por [ARAMEO.NET](#)

## ARAMEO.NET

**Editorial Mar Caribe**

**Metodología de la investigación científica  
aplicada a la inteligencia artificial y la ciencia de  
datos: Enfoque general**

**Colonia, Uruguay**

**2025**

# Índice

Introducción.....	7
Capítulo 1.....	10
El Imperativo Metodológico en la Era del Algoritmo .....	10
1. El Cisma Epistemológico: ¿Ciencia o Ingeniería? .....	11
1.1 La Distinción Teleológica y Procedimental .....	11
Tabla 1: Contrastación Metodológica entre Ciencia e Ingeniería en el Contexto de la IA .....	12
1.2 El Debate de la "Alquimia" y la Madurez del Campo .....	13
2. La Crisis de Reproducibilidad: Anatomía de un Fallo Sistémico .....	14
2.1 Magnitud Cuantitativa del Problema .....	15
2.2 Dimensiones de la Reproducibilidad en IA .....	15
2.3 El Factor Oculto: Sensibilidad a la Semilla Aleatoria .....	16
2.4 La Respuesta Institucional: Listas de Verificación y Código Obligatorio .....	17
3. Formulación del Problema y Diseño Experimental.....	18
3.1 De la Optimización a la Hipótesis .....	18
3.2 El Control de Variables y los Estudios de Ablación.....	19
4. El Sustrato de los Datos: Documentación y Gobernanza.....	19
4.1 "Garbage In, Garbage Out" y la Fuga de Datos.....	20
4.2 Datasheets for Datasets: Estandarización de la Transparencia.....	20
5. Métricas y Evaluación: Más Allá de la Precisión.....	21
5.1 Taxonomía de Métricas y sus Implicaciones Metodológicas.....	21
Tabla 2: Análisis Metodológico de Métricas de Evaluación .....	22
5.2 La Ley de Goodhart y el "Hacking" de Métricas .....	23
5.3 Significancia Estadística .....	24
6. Ética como Componente Metodológico .....	24
6.1 Auditoría de Sesgos y Evaluación Desagregada.....	24
6.2 Model Cards: Documentación del Artefacto .....	25
7. Instituciones y Reformas: Hacia una Ciencia Abierta y Pre-registrada .....	25

7.1 El Problema de la Revisión por Pares y la "Mathiness" .....	25
7.2 Pre-registro (Preregistration) y Reportes Registrados.....	26
Capítulo 2. ....	27
La Reconfiguración Epistemológica .....	27
1. La Evolución de los Paradigmas Científicos y el Diluvio de Datos .....	28
1.1 El Surgimiento del Cuarto Paradigma.....	28
1.2 El Debate sobre el "Fin de la Teoría" .....	29
1.3 Razonamiento Abductivo: La Síntesis Metodológica.....	30
Tabla 3: Síntesis metodológica.....	30
2. Arquitecturas de Procesos y Ciclos de Vida en la Investigación .....	31
2.1 El Marco CRISP-DM: Estándar y Adaptación.....	31
Tabla 4: El Marco CRISP-DM .....	32
2.2 Alternativas Académicas: KDD y SEMMA .....	33
2.3 Hacia un Modelo Híbrido: CRISP-DS y Agile .....	33
3. La Crisis de Rigor y Reproducibilidad .....	34
3.1 Anatomía del Fracaso: Fuga de Datos (Data Leakage).....	34
3.2 Listas de Verificación de Reproducibilidad (Checklists).....	35
4. Documentación como Metodología Científica .....	36
4.1 Datasheets for Datasets (Hojas de Datos para Conjuntos de Datos).....	36
4.2 Model Cards for Model Reporting (Tarjetas de Modelo) .....	37
5. La IA Co-Científica y Métodos Generativos .....	38
5.1 Generación Automatizada de Hipótesis.....	38
5.2 Riesgos de los LLMs en la Investigación .....	39
6. Ética, Gobernanza y el Nuevo Marco Regulatorio .....	39
6.1 La Declaración de Impacto Amplio (Broader Impact Statement).....	40
6.2 Marcos Regulatorios como Restricciones Metodológicas.....	40
6.3 Ciclos de Vida de IA Responsable.....	41
Capítulo 3.....	43
Prolegómenos: La Crisis Epistemológica del Cuarto Paradigma.....	43
1. Rigor metodológico de la inteligencia artificial.....	43

1.1. La Evolución de la Indagación Científica .....	44
Tabla 5: Paradigmas de Jim Gray .....	44
1.2. La Falsa Equivalencia: Ingeniería vs. Investigación .....	45
Tabla 6: Divergencias Metodológicas Fundamentales <sup>8</sup> .....	46
2. La Anatomía de la Crisis de Reproducibilidad .....	47
2.1. Los Pecados Capitales de la Investigación en IA .....	48
2.2. La Respuesta Institucional: Listas de Verificación y Retracciones .....	49
3. El Marco Metodológico Propuesto: El Modelo Anidado y el Diseño Experimental .....	50
3.1. Las Cuatro Capas de Validación .....	50
3.2. Diseño de Experimentos (DoE) en el Silicio .....	51
Tabla 7: Estrategias de Validación Experimental <sup>19</sup> .....	52
4. Estructura de un Plan de Estudios de Posgrado en Investigación con IA (ejes temáticos) .....	53
Semestre I: Fundamentos y Filosofía .....	53
Semestre II: Diseño y Preparación .....	53
Semestre III: Experimentación y Modelado (El Núcleo Técnico) .....	54
Semestre IV: Análisis y Comunicación .....	54
5. El Factor Humano: Consejos para el Investigador Novel .....	54
5.1. La Elección del Problema y la Ventaja Competitiva .....	54
5.2. La Responsabilidad Ética y el Uso de IA Generativa .....	55
Capítulo 4 .....	56
La epistemología de lo artificial: un marco integral para la metodología de la investigación científica en inteligencia artificial y ciencia de datos .....	56
1. La convergencia entre computación y empirismo .....	56
1.1 La tensión entre predicción y explicación .....	56
1.2 El estado epistémico de la investigación basada en datos .....	57
2. Modelos de procesos estructurados: desde los estándares industriales hasta el rigor científico .....	58
2.1 Crítica de los modelos industriales estándar .....	58
2.2 El proceso de Ciencia de Datos en Equipo (TDSP) y adaptaciones modernas	

.....	59
2.3 Hacia un modelo de procesos científicos: GCRISP-DS y más allá .....	59
2.4 Análisis comparativo de marcos metodológicos.....	59
Tabla 8: Características de métodos de investigación tradicionales y con IA ....	60
3. Generación de hipótesis y sesgo inductivo .....	60
3.1 Sesgo inductivo como hipótesis .....	60
3.2 IA generativa y generación automatizada de hipótesis .....	61
3.3 El ciclo de hipótesis basadas en datos .....	61
4. Diseño experimental en el laboratorio computacional .....	62
4.1 La necesidad de líneas base.....	62
4.2 Variables de control y estudios de ablación.....	63
4.3 Fuga de datos: El factor de confusión de la validez .....	63
4.4 Datos sintéticos como control experimental .....	64
5. Rigor estadístico y evaluación del modelo.....	64
5.1 Pruebas de hipótesis para la comparación de modelos.....	64
5.2 Intervalos de confianza (IC) .....	65
5.3 Pruebas de Esfuerzo Adversariales .....	65
Tabla 9: Métodos de evaluación estadística .....	66
6. Reproducibilidad, transparencia y pre-registro .....	66
6.1 La Lista de Verificación de Reproducibilidad.....	67
6.2 Pre-registro de la investigación .....	67
6.3 Control de versiones para datos (DVC).....	67
7. Causalidad y explicación: más allá de la correlación .....	68
7.1 Inferencia causal vs. modelado predictivo .....	68
7.2 IA explicable (XAI) como falsificación científica.....	68
Tabla 10: Recomendaciones metodológicas clave para investigadores en IA ....	69
Conclusión .....	71
Bibliografía.....	73

# Introducción

La sinergia entre la ciencia de datos, la inteligencia artificial (IA) y la investigación académica ha precipitado una transformación radical en los modos de producción de conocimiento. Nos encontramos ante una disciplina que, aunque heredera de la estadística clásica y las ciencias de la computación, ha evolucionado hacia una entidad autónoma con dinámicas propias, desafíos inéditos y, críticamente, una fragilidad metodológica alarmante. La premisa fundamental que motiva la elaboración del libro *"Metodología de la investigación científica aplicada a la inteligencia artificial y la ciencia de datos"* no es la carencia de avances técnicos —los cuales se producen a un ritmo vertiginoso— sino la ausencia de un marco normativo unificado que garantice que dichos avances sean robustos, reproducibles y éticamente sostenibles.

La investigación en IA ha transitado de ser una curiosidad académica a convertirse en el motor de infraestructuras críticas en salud, finanzas y gobernanza. Sin embargo, este despliegue ha ocurrido a menudo bajo una lógica de "ingeniería de software" —donde lo importante es que el sistema funcione— en detrimento de la lógica científica, que exige comprender *por qué* funciona y bajo qué condiciones dejaría de hacerlo. Esta tensión entre la predicción empírica y la explicación teórica ha generado lo que algunos autores denominan una "crisis de adolescencia" de la disciplina, caracterizada por una confusión terminológica generalizada, prácticas de evaluación inconsistentes y una deuda técnica y ética que amenaza con socavar la confianza pública en los sistemas inteligentes.

El presente libro despliega un análisis exhaustivo para fundamentar el planteamiento del problema y definir el alcance de esta obra propuesta. A través de una revisión profunda del estado del arte, se evidencia que la comunidad científica se



enfrenta a barreras procedimentales y técnicas que van desde la gestión de semillas aleatorias en redes neuronales profundas hasta la contaminación de datos en grandes modelos de lenguaje (LLMs). La obra propuesta busca llenar este vacío, no como un manual más de programación en Python o R, sino como un tratado sobre el diseño, ejecución y comunicación de la ciencia en la era de los algoritmos de caja negra. Se postula aquí una integración de las "Ciencias de lo Artificial" de Herbert Simon con las metodologías modernas de *Design Science Research* (DSR), estableciendo un puente necesario entre la creación de artefactos tecnológicos y la generación de teoría científica válida.

El problema central que este libro aborda es la desconexión sistémica entre la capacidad técnica de modelado y el rigor metodológico necesario para validar dichos modelos como conocimiento científico. Esta desconexión no es un fallo puntual, sino una patología estructural que se manifiesta en múltiples dimensiones, desde la formulación de hipótesis hasta la revisión por pares. A continuación, se diseccionan las facetas críticas de esta problemática.

La reproducibilidad es el *sine qua non* del método científico; sin ella, la acumulación de conocimiento es imposible. Sin embargo, la investigación en IA y aprendizaje automático (Machine Learning - ML) enfrenta una crisis severa en este aspecto. Investigaciones recientes sugieren que una fracción significativa de los resultados publicados en conferencias de primer nivel (NeurIPS, ICML, AAAI) no pueden ser replicados por terceros, e incluso presentan dificultades para los propios autores originales tiempo después de la publicación.

En síntesis, se discierne sobre la fricción fundamental entre la estadística tradicional y la ciencia de datos moderna. Históricamente, el método científico se ha basado en la formulación de hipótesis *a priori* y el uso de modelos estadísticos parsimoniosos para inferir relaciones causales o correlacionales en una población. La

ciencia de datos, por el contrario, ha priorizado la capacidad predictiva sobre datos no vistos, a menudo sacrificando la interpretabilidad y la parsimonia en favor de modelos de alta complejidad.

# Capítulo 1.

## El Imperativo Metodológico en la Era del Algoritmo

La inteligencia artificial (IA) y la ciencia de datos han transitado, en un lapso sorprendentemente breve, de ser curiosidades académicas y subdisciplinas de las ciencias de la computación a convertirse en los motores fundamentales de la transformación industrial, social y científica del siglo XXI. Sin embargo, este ascenso vertiginoso ha traído consigo una crisis de identidad epistemológica. A medida que los modelos de aprendizaje automático (Machine Learning, ML) y aprendizaje profundo (Deep Learning, DL) alcanzan capacidades sobrehumanas en tareas específicas —desde el diagnóstico médico hasta la generación de lenguaje natural—, la comunidad científica se enfrenta a una paradoja inquietante: nuestra capacidad para *construir* sistemas inteligentes ha superado nuestra capacidad para *entenderlos* y *evaluarlos* con rigor científico.<sup>1</sup>

Este libro, titulado "Metodología de investigación científica aplicada a la inteligencia artificial y la ciencia de datos: Enfoque general", nace de la necesidad urgente de cerrar la brecha entre la ingeniería de software pragmática y la investigación científica rigurosa. La literatura actual sugiere que gran parte de la producción académica en IA, a pesar de su volumen y velocidad, adolece de deficiencias metodológicas graves que amenazan la validez de sus hallazgos, la reproducibilidad de sus experimentos y la seguridad de sus aplicaciones.<sup>2,3</sup> Nos encontramos en un momento que algunos teóricos han calificado como una era de "alquimia", donde la intuición y el ensayo y error predominan sobre la teoría

fundamentada y el diseño experimental controlado.<sup>4,5</sup>

El propósito de esta obra es sistematizar los fundamentos de la investigación científica aplicados específicamente al contexto estocástico, basado en datos y computacionalmente intensivo de la IA moderna. No se trata simplemente de aplicar el método científico tradicional, sino de adaptarlo a un dominio donde el "objeto de estudio" es un artefacto digital que cambia dinámicamente, y donde los datos actúan simultáneamente como el instrumento de observación y el sujeto del experimento.<sup>6</sup> A través de un análisis exhaustivo de la literatura reciente, las directrices de las principales conferencias (NeurIPS, ICML, ICLR) y los debates teóricos más relevantes, esta introducción establece el marco conceptual para una nueva praxis científica en la inteligencia artificial.

## **1. El Cisma Epistemológico: ¿Ciencia o Ingeniería?**

Para abordar la metodología de investigación en IA, primero debemos desentrañar la tensión fundamental que subyace en la disciplina: la dicotomía entre el enfoque científico y el enfoque de ingeniería. Aunque ambos comparten herramientas y entornos, sus objetivos teleológicos son divergentes, y la confusión entre ambos es la fuente de numerosos errores metodológicos en la literatura contemporánea.

### **1.1 La Distinción Teleológica y Procedimental**

La ciencia, en su definición clásica, busca la generación de conocimiento a través de la comprensión de fenómenos naturales o artificiales. El método científico es un proceso recursivo de análisis y síntesis destinado a reducir la incertidumbre sobre cómo funciona el mundo. Sus fases —observación, formulación de hipótesis, experimentación y análisis— están diseñadas para validar o refutar teorías

explicativas.<sup>6</sup> En este contexto, un resultado negativo (la refutación de una hipótesis) es tan valioso como uno positivo, pues ambos aportan información sobre la realidad subyacente.<sup>2,7</sup>

Por el contrario, la ingeniería se centra en la utilidad. El proceso de diseño en ingeniería busca la creación de soluciones a problemas específicos bajo restricciones dadas. Su ciclo implica definir requisitos, prototipar, probar y rediseñar hasta alcanzar un desempeño satisfactorio.<sup>8</sup> El éxito se mide por la funcionalidad del artefacto, no necesariamente por la comprensión profunda de sus principios operativos.

El problema actual en la investigación de IA es que gran parte de los trabajos publicados siguen estructuralmente un proceso de ingeniería (optimizar una métrica de rendimiento en un benchmark) pero se presentan retóricamente como ciencia (descubrir propiedades fundamentales del aprendizaje). Esto conduce a lo que Lipton y Steinhardt han denominado "tendencias preocupantes" en la erudición del ML, donde la explicación se confunde con la especulación y las ganancias empíricas se atribuyen erróneamente a novedades teóricas sin la debida evidencia causal (ver Tabla 1).<sup>4,9</sup>

**Tabla 1: Contrastación Metodológica entre Ciencia e Ingeniería en el Contexto de la IA**

Dimensión	Método Científico	Enfoque de Ingeniería/Diseño	Práctica Común en Investigación de IA (El Problema)
Objetivo Primario	Comprensión / Explicación (¿Por qué)	Solución / Producto (¿Funciona?)	Estado del Arte (SOTA) / Maximización de

	funciona?)		Métricas
<b>Punto de Partida</b>	Pregunta de investigación / Hipótesis falsable	Definición del problema / Requisitos del usuario	Conjunto de datos de referencia (Benchmark)
<b>Rol de los Datos</b>	Evidencia para probar la hipótesis	Insumo para entrenar la solución	El terreno de competencia
<b>Valor del Fallo</b>	Alto (Refutación de teoría, avance del conocimiento)	Bajo (El prototipo no sirve, se descarta)	Nulo (Resultados no publicables, sesgo de publicación)
<b>Validación</b>	Reproducibilidad y control de variables	Pruebas de aceptación y rendimiento en producción	Superar al modelo anterior en la tabla de clasificación

## 1.2 El Debate de la "Alquimia" y la Madurez del Campo

Esta tensión alcanzó su punto álgido en el famoso debate sobre la "alquimia" en el aprendizaje automático. Ali Rahimi, en su discurso de aceptación del premio "Test of Time" en NIPS 2017, argumentó que el aprendizaje automático se había convertido en una forma de alquimia moderna.<sup>5,10</sup> Al igual que los alquimistas medievales descubrieron técnicas útiles (metalurgia, tintes) mediante el ensayo y error sin una teoría química subyacente, los investigadores de IA actuales combinan arquitecturas de redes neuronales, funciones de activación y optimizadores basándose en la intuición y la "sabiduría popular" más que en una comprensión teórica rigurosa.<sup>5</sup>

Yann LeCun y otros defensores del enfoque empírico contraargumentaron que la ingeniería a menudo precede a la ciencia teórica (la máquina de vapor existió antes

que la termodinámica).<sup>11,12</sup> Sin embargo, la crítica de la alquimia resalta riesgos metodológicos graves:

- **Fragilidad de los Sistemas:** Sin entender los mecanismos causales, es difícil predecir cuándo y cómo fallarán los modelos, especialmente ante datos fuera de la distribución (OOD).<sup>5</sup>
- **Opacidad y Falta de Interpretabilidad:** La complejidad de los modelos de "caja negra" impide distinguir si un sistema funciona por razones legítimas o porque ha aprendido correlaciones espurias (como detectar un barco solo si hay agua alrededor).<sup>5</sup>
- **Deuda Técnica:** La acumulación de trucos y "parches" algorítmicos sin fundamento teórico crea sistemas difíciles de mantener y mejorar sistemáticamente.

La transición de la IA hacia una ciencia madura requiere, por tanto, adoptar metodologías que prioricen la explicabilidad, la hipótesis falsable y la validación robusta por encima de la mera optimización de métricas.<sup>1</sup>

## 2. La Crisis de Reproducibilidad: Anatomía de un Fallo Sistémico

La piedra angular de la ciencia es la reproducibilidad. Si un hallazgo no puede ser replicado por un investigador independiente bajo condiciones similares, no constituye conocimiento científico válido. En la última década, la ciencia en general ha enfrentado una "crisis de reproducibilidad", y la IA no ha sido inmune; de hecho, en muchos aspectos, es el epicentro de una nueva variante de esta crisis.<sup>3</sup>

## 2.1 Magnitud Cuantitativa del Problema

Estudios sistemáticos recientes arrojan cifras alarmantes. Un estudio de replicación de 2024, que analizó 30 estudios de IA altamente citados, encontró que solo el **50%** de los artículos pudieron ser reproducidos total o parcialmente.<sup>3</sup> Aún más revelador es el análisis de los factores que determinan el éxito:

- La disponibilidad de código y datos es el predictor más fuerte de reproducibilidad. El **86%** de los artículos que compartían ambos pudieron ser reproducidos.
- Cuando solo se compartían los datos (sin el código), la tasa de éxito caía dramáticamente al 33%.<sup>3,13</sup>
- Sorprendentemente, la *calidad* de la documentación del código (si está limpio, comentado o versionado) resultó ser menos crítica que su mera *existencia*. El código, incluso si es desordenado, actúa como la especificación exacta del experimento, capturando detalles que las descripciones en lenguaje natural del "paper" omiten.<sup>3,14</sup>

## 2.2 Dimensiones de la Reproducibilidad en IA

Para abordar esta crisis, es necesario desambiguar el término "reproducibilidad", que a menudo se usa de manera laxa.<sup>15,16</sup> Un marco reciente propone una taxonomía clara para estandarizar los esfuerzos de validación<sup>17</sup>:

1. **Repetibilidad (Repeatability):** El mismo equipo, usando el mismo código y entorno experimental, puede obtener el mismo resultado. Esto verifica la estabilidad básica del código.
2. **Reproducibilidad (Reproducibility):** Un equipo *diferente*, usando el mismo código y datos originales, obtiene el mismo resultado. Esto verifica que los



hallazgos no dependen del artífice original.

3. **Replicabilidad (Replicability):** Un equipo diferente, escribiendo su *propio* código basado en la descripción del método y recolectando nuevos datos, llega a conclusiones científicas similares. Esto valida la robustez de la teoría o el método más allá de una implementación específica.

La mayoría de los fallos actuales ocurren en los niveles de reproducibilidad y replicabilidad, impulsados por la falta de transparencia y la sensibilidad extrema de los algoritmos de optimización no convexa.

## 2.3 El Factor Oculto: Sensibilidad a la Semilla Aleatoria

Uno de los hallazgos metodológicos más inquietantes es la extrema sensibilidad de los modelos de aprendizaje profundo a la inicialización aleatoria (random seeds). Las redes neuronales se inicializan con pesos aleatorios, y el orden en que se presentan los datos durante el entrenamiento (Stochastic Gradient Descent) también es aleatorio.

Investigaciones han demostrado que cambiar *únicamente* la semilla aleatoria puede alterar drásticamente el rendimiento del modelo. En un caso documentado con la arquitectura LeNet5, la precisión varió entre 8.6% y 99.0% a través de 16 ejecuciones idénticas, una diferencia del 90.4% atribuible puramente al azar.<sup>18</sup> Aunque este es un caso extremo, en modelos modernos de Procesamiento de Lenguaje Natural (NLP) como BERT o RoBERTa, la varianza inducida por la semilla es a menudo mayor que la mejora marginal que los investigadores reportan como "nuevo estado del arte".<sup>18,19</sup>

Esta realidad ha fomentado una mala práctica conocida como "seed hacking" o "p-hacking computacional": los investigadores ejecutan su modelo con docenas de semillas diferentes y reportan solo la mejor ejecución, presentándola como

representativa del método.<sup>20,21</sup> Una metodología rigurosa exige, por el contrario, reportar la media y la desviación estándar de múltiples ejecuciones ( $k \geq 5$  o 10) y realizar pruebas de significancia estadística para confirmar que la mejora no es ruido estocástico.<sup>22,23</sup>

## 2.4 La Respuesta Institucional: Listas de Verificación y Código Obligatorio

Ante la evidencia de esta crisis, las principales conferencias (NeurIPS, ICML) han implementado intervenciones estructurales. La introducción de la Lista de Verificación de Reproducibilidad de NeurIPS (Reproducibility Checklist) marcó un hito en la gobernanza de la investigación.<sup>25</sup> Este instrumento obliga a los autores a declarar explícitamente:

- Si se incluye el código y los datos.
- La especificación de las dependencias y el entorno de computación.
- El número de ejecuciones y las semillas utilizadas.
- Las barras de error y las pruebas de significancia estadística.

Los análisis de impacto muestran que estas políticas funcionan. La disponibilidad de código en NeurIPS aumentó de menos del 50% a casi el 75% tras la implementación de estas medidas.<sup>25</sup> Además, se observó que los revisores asignan puntuaciones más altas a los trabajos que incluyen código y listas de verificación completas, creando un incentivo virtuoso hacia la transparencia.<sup>25</sup> Sin embargo, persisten desafíos, especialmente con los Grandes Modelos de Lenguaje (LLMs) propietarios, donde la falta de acceso a los datos de entrenamiento y al código fuente hace que la investigación sea esencialmente irreproducible por diseño, un problema que amenaza con bifurcar el campo entre la "ciencia abierta" académica y la "ciencia

cerrada" corporativa.<sup>3,24</sup>

### 3. Formulación del Problema y Diseño Experimental

La metodología científica comienza mucho antes de escribir la primera línea de código; comienza con la formulación del problema. En la IA, la transición de una mentalidad de ingeniería a una científica requiere redefinir cómo se plantean las preguntas de investigación.

#### 3.1 De la Optimización a la Hipótesis

Tradicionalmente, un "paper" de IA se estructura en torno a una afirmación de rendimiento: *"Proponemos la arquitectura X que mejora la precisión en la tarea Y en un Z%"*. Esto es una formulación de ingeniería. Una formulación científica, en cambio, debe basarse en una hipótesis explicativa: *"Hipótesis: La incorporación del mecanismo de atención A permite al modelo capturar dependencias de largo alcance que son invisibles para la convolución B, lo cual debería resultar en una mejora específica en subconjuntos de datos con alta complejidad sintáctica"*.<sup>2,25,26</sup>

Esta distinción es crucial porque dicta el diseño experimental. Si el objetivo es solo optimizar, cualquier "truco" es válido. Si el objetivo es entender, el diseño debe aislar variables<sup>27,28</sup>. El uso de Métodos Formales para especificar propiedades deseadas (como robustez o seguridad) antes de la implementación empírica es una práctica emergente que añade rigor matemático a la formulación del problema, permitiendo verificar no solo si el modelo acierta, sino si cumple con garantías lógicas predefinidas.<sup>29</sup>

### 3.2 El Control de Variables y los Estudios de Ablación

Un error metodológico omnipresente en la literatura de IA es la falta de controles adecuados, lo que lleva a identificar erróneamente la fuente de las mejoras empíricas.<sup>4,30</sup> A menudo, los investigadores proponen un modelo complejo con múltiples componentes nuevos y lo comparan con una línea base (baseline) débil o mal ajustada.

Para combatir esto, la metodología rigurosa exige:

1. **Líneas Base Fuertes:** Los modelos de comparación deben recibir el mismo nivel de ajuste de hiperparámetros (hyperparameter tuning) que el modelo propuesto.  
<sup>31</sup> Se ha demostrado repetidamente que modelos simples (como regresión logística o LSTMs básicos) pueden igualar a modelos complejos "SOTA" si se ajustan correctamente.<sup>10</sup>
2. **Estudios de Ablación (Ablation Studies):** Si un nuevo modelo introduce tres cambios (A, B y C), el investigador debe evaluar el modelo con todas las combinaciones posibles (solo A, A+B, B+C, etc.) para determinar la contribución marginal de cada componente. Sin ablación, es imposible saber si la mejora se debe a la innovadora arquitectura neuronal o simplemente a una mejor técnica de preprocesamiento de datos.<sup>2, 32</sup>

## 4. El Sustrato de los Datos: Documentación y Gobernanza

En la ciencia impulsada por datos, el conjunto de datos (dataset) no es un mero insumo pasivo; es el instrumento de observación y el árbitro de la verdad. Un conjunto de datos sesgado, ruidoso o mal documentado invalida cualquier conclusión extraída

de él, independientemente de la sofisticación del algoritmo.<sup>33,34</sup> Por ello, la gestión de datos debe elevarse a la categoría de disciplina metodológica primaria.

#### **4.1 "Garbage In, Garbage Out" y la Fuga de Datos**

El fenómeno de "Basura entra, basura sale" es bien conocido, pero en la investigación moderna se manifiesta de formas sutiles y perniciosas, principalmente a través de la fuga de datos (data leakage). La fuga ocurre cuando información del conjunto de prueba (test set) se filtra inadvertidamente en el conjunto de entrenamiento, permitiendo al modelo "hacer trampa" y memorizar respuestas en lugar de generalizar patrones.<sup>14</sup>

Un análisis exhaustivo de Kapoor y Narayanan reveló que la fuga de datos es responsable de una gran cantidad de resultados irreproducibles en la literatura, donde modelos complejos parecían superar a los tradicionales solo porque, por ejemplo, se realizaba la selección de características o la normalización de datos *antes* de separar los conjuntos de entrenamiento y prueba.<sup>14,35</sup> La metodología estricta dicta que el conjunto de prueba debe mantenerse en una "bóveda", inaccesible para cualquier proceso de decisión, ajuste o preprocesamiento hasta la evaluación final.

#### **4.2 Datasheets for Datasets: Estandarización de la Transparencia**

Para mitigar los problemas de calidad y sesgo en los datos, la comunidad ha adoptado el marco de "Datasheets for Datasets" (Hojas de Datos para Conjuntos de Datos), propuesto por Gebru et al..<sup>32,36</sup> Inspirado en la industria electrónica, donde cada componente viene con una hoja de especificaciones detallada, este marco exige que cada conjunto de datos público vaya acompañado de documentación que responda preguntas críticas:

- **Motivación:** ¿Para qué propósito fue creado el conjunto de datos?
- **Composición:** ¿Qué representan las instancias? ¿Hay subgrupos demográficos? ¿Hay datos confidenciales?
- **Proceso de Recolección:** ¿Cómo se obtuvieron los datos? ¿Se obtuvo consentimiento ético?
- **Preprocesamiento:** ¿Qué limpieza, filtrado o aumento de datos se realizó?
- **Usos Recomendados y No Recomendados:** ¿Para qué tareas es adecuado y para cuáles no?

Esta documentación permite realizar auditorías de datos y ayuda a los investigadores a evaluar si un conjunto de datos es apropiado para su hipótesis específica, evitando el uso ciego de benchmarks que pueden contener sesgos ocultos.

34,37

## 5. Métricas y Evaluación: Más Allá de la Precisión

La elección de la métrica de evaluación es una decisión epistemológica que define qué constituye el "éxito". En la práctica actual, existe una dependencia excesiva de métricas simplistas como la precisión global (accuracy), que a menudo ocultan fallos críticos del modelo.

### 5.1 Taxonomía de Métricas y sus Implicaciones Metodológicas

Una metodología rigurosa requiere seleccionar métricas alineadas con la naturaleza del problema y los costos de los errores. A continuación se presenta una síntesis de las métricas clave y sus contextos de aplicación apropiados (ver Tabla 2).<sup>36</sup>

**Tabla 2: Análisis Metodológico de Métricas de Evaluación**

<b>Categoría</b>	<b>Métrica</b>	<b>Ventaja Metodológica</b>	<b>Riesgo/Limitación (El "Pecado" Metodológico)</b>	<b>Contexto Ideal</b>
<b>Clasificación</b>	<b>Accuracy (Exactitud)</b>	Intuitiva y fácil de comunicar.	Engañosa en datos desbalanceados (Paradoja de la exactitud).	Clases perfectamente balanceadas.
	<b>Precision &amp; Recall</b>	Desglosa el tipo de error (Falsos Positivos vs Falsos Negativos).	Requiere analizar dos números o elegir un compromiso.	Problemas de "aguja en un pajar" (fraude, diagnóstico).
	<b>F1 Score</b>	Media armónica, penaliza valores extremos de P o R.	Puede ocultar el compromiso específico P/R.	Comparación general en datos desbalanceados.
	<b>AUC-ROC</b>	Independiente del umbral de decisión. Mide discriminabilidad pura.	Puede ser optimista si los Falsos Positivos son muy raros.	Evaluar la calidad del modelo antes de definir el punto de operación.
<b>Regresión</b>	<b>MSE / RMSE</b>	Penaliza fuertemente los grandes errores (cuadrático).	Muy sensible a valores atípicos (outliers).	Cuando un error grande es catastrófico.

	<b>MAE</b>	Robusto a outliers, interpretación directa en unidades de la variable.	Gradientes constantes dificultan optimización fina.	Cuando el costo del error es lineal.
<b>Clustering</b>	<b>Silhouette Score</b>	Mide cohesión y separación sin necesitar etiquetas (intrínseco).	Costoso computacionalmente $O(N^2)$ .	Validación de estructura interna de datos.

## 5.2 La Ley de Goodhart y el "Hacking" de Métricas

Un fenómeno crítico en la evaluación de IA es la Ley de Goodhart: *"Cuando una medida se convierte en un objetivo, deja de ser una buena medida"*.<sup>40</sup> En el aprendizaje automático, esto se manifiesta como Reward Hacking (hacking de recompensa). Si entrenamos un agente para "minimizar errores en un videojuego", el agente podría aprender a pausar el juego indefinidamente para evitar perder, cumpliendo la métrica pero violando el espíritu de la tarea.<sup>38,42</sup>

Para contrarrestar la Ley de Goodhart, la metodología de investigación debe incluir:

- **Métricas Proxy vs. Métricas Reales:** Reconocer explícitamente que la función de pérdida es solo un proxy del objetivo real.<sup>39</sup>
- **Evaluación Multidimensional:** Nunca optimizar una sola métrica. Reportar un conjunto de indicadores (trade-offs).<sup>40</sup>
- **Evaluación Humana:** En tareas generativas (traducción, resumen), las métricas automáticas (BLEU, ROUGE) son insuficientes y a menudo no correlacionan con la calidad percibida.<sup>41,42</sup> La validación humana sigue siendo el "gold standard".<sup>43</sup>



## 5.3 Significancia Estadística

Finalmente, es imperativo reintroducir las pruebas de hipótesis estadística en la comparación de modelos. Afirmar que el Modelo A es mejor que el Modelo B porque su exactitud es 94.5% vs 94.2% carece de rigor si no se acompaña de una prueba estadística (como el Test de McNemar para clasificación o t-tests para regresión) que demuestre que la diferencia no es producto del azar.<sup>23</sup> La ausencia de estas pruebas es una de las causas principales de la proliferación de resultados espurios en la literatura.<sup>15</sup>

## 6. Ética como Componente Metodológico

Históricamente, la ética se consideraba un control externo a la investigación científica (comités de ética). En la IA, la ética es intrínseca a la metodología técnica.<sup>44</sup> Un modelo sesgado no es solo "injusto"; es científicamente defectuoso porque ha fallado en aprender la verdadera distribución invariante del fenómeno y se ha sobreajustado a correlaciones espurias o prejuicios históricos presentes en los datos de entrenamiento.<sup>45</sup>

### 6.1 Auditoría de Sesgos y Evaluación Desagregada

La metodología estándar de reportar una única cifra de rendimiento global (e.g., "90% de precisión") oculta fallos sistemáticos en subgrupos minoritarios. Esto se conoce como sesgo de agregación. Una investigación rigurosa debe emplear evaluación desagregada: reportar el rendimiento del modelo desglosado por variables sensibles (género, raza, edad, ubicación geográfica).<sup>46,47</sup>

Por ejemplo, un sistema de reconocimiento facial puede tener una precisión

global del 99%, pero una precisión del 60% en mujeres de piel oscura. Sin evaluación desagregada, este fallo catastrófico permanece invisible metodológicamente hasta que el sistema causa daño en el mundo real.<sup>45,48</sup>

## 6.2 Model Cards: Documentación del Artefacto

Así como los *Datasheets* documentan los datos, los Model Cards (Tarjetas de Modelo), propuestos por Mitchell et al., documentan el modelo resultante.<sup>47,49</sup> Un Model Card es un requisito metodológico para la publicación responsable que detalla:

- **Uso Previsto:** ¿Para qué fue diseñado el modelo?
- **Limitaciones y Fuera de Alcance:** ¿En qué contextos NO debe usarse el modelo? (e.g., "No usar para diagnóstico clínico autónomo").
- **Métricas de Equidad:** Resultados de la evaluación desagregada.
- **Factores Ambientales:** Impacto de carbono del entrenamiento.<sup>50</sup>

La adopción de Model Cards transforma el modelo de una "caja negra" mágica a una herramienta de ingeniería con especificaciones de seguridad definidas, permitiendo a usuarios y reguladores evaluar su idoneidad y riesgo.<sup>51</sup>

## 7. Instituciones y Reformas: Hacia una Ciencia Abierta y Pre-registrada

La metodología no existe en el vacío; es moldeada por las instituciones que publican y financian la investigación. Las conferencias de IA (NeurIPS, ICML, ICLR) actúan como guardianes (gatekeepers) de la calidad científica.

### 7.1 El Problema de la Revisión por Pares y la "Mathiness"

El crecimiento exponencial de envíos a conferencias (miles de papers por evento) ha saturado el sistema de revisión por pares, resultando en revisiones ruidosas y de calidad variable.<sup>52</sup> Esto incentiva la "mathiness": el uso de matemáticas densas y complejas para impresionar a los revisores y ocultar la falta de sustancia empírica o teórica.<sup>4,52</sup> Los revisores, a menudo abrumados y sin experiencia profunda en el subnicho específico, pueden asumir que las ecuaciones son correctas en lugar de verificarlas, un fenómeno que erosiona la confianza en la literatura.

## 7.2 Pre-registro (Preregistration) y Reportes Registrados

Para combatir el sesgo de publicación y el "HARKing" (Hipocresía de formular la hipótesis después de conocer los resultados), la comunidad de IA está comenzando a experimentar con Reportes Registrados (Registered Reports).<sup>53,54</sup>

En este modelo, los investigadores envían su introducción y metodología antes de realizar los experimentos. Si el diseño es sólido y la pregunta relevante, el paper es "aceptado en principio", independientemente de si los resultados son positivos o negativos. Esto tiene ventajas profundas:

- **Elimina el incentivo de "hackear" resultados:** Ya no es necesario obtener un "SOTA" para publicar.
- **Valoriza los resultados negativos:** Saber que un método prometedor *no* funciona es tan valioso como saber que sí.<sup>55</sup>
- **Mejora el diseño:** La revisión por pares ocurre cuando todavía se puede corregir la metodología, no cuando el trabajo ya está hecho.<sup>56,57</sup>

Aunque su adopción en IA es incipiente en comparación con la psicología o la medicina, su implementación en talleres y tracks especiales de conferencias señala un cambio cultural hacia la integridad metodológica sobre la novedad sensacionalista.<sup>58</sup>

## Capítulo 2.

# La Reconfiguración Epistemológica

La historia de la ciencia es, en gran medida, la historia de las herramientas que utilizamos para observar el universo. Desde el telescopio de Galileo hasta el microscopio electrónico, cada avance instrumental ha redefinido los límites de lo observable y, por ende, de lo conocible. En el siglo XXI, el instrumento científico por excelencia ha dejado de ser puramente físico para convertirse en computacional y algorítmico. La Inteligencia Artificial (IA) y la Ciencia de Datos han trascendido su rol inicial como objetos de estudio para convertirse en los mecanismos mismos de la indagación científica. Esta transición representa una transformación fundamental en el método científico, desafiando las normas epistemológicas establecidas y ofreciendo capacidades sin precedentes para comprender fenómenos de complejidad irreductible.

Al situarnos en el precipicio de lo que Jim Gray denominó el "Cuarto Paradigma" de la ciencia —el descubrimiento intensivo de datos—, resulta imperativo formalizar los métodos mediante los cuales extraemos conocimiento veraz de los exabytes de información generados diariamente.<sup>1</sup> Este informe delinea el enfoque general de la metodología de investigación científica en IA, trazando el terreno desde la generación de hipótesis asistida por máquinas hasta el despliegue de modelos y la evaluación de su impacto social bajo nuevos marcos regulatorios.

# 1. La Evolución de los Paradigmas Científicos y el Diluvio de Datos

## 1.1 El Surgimiento del Cuarto Paradigma

Durante siglos, el método científico osciló pendularmente entre dos modos primarios: el empírico, dedicado a la descripción de fenómenos naturales (como la taxonomía de Linneo), y el teórico, enfocado en el uso de modelos y generalizaciones matemáticas (como las leyes de Newton) para predecir comportamientos.<sup>59</sup> A mediados del siglo XX, la llegada de los ordenadores digitales introdujo un tercer paradigma: la ciencia computacional, que permitió simular fenómenos complejos — desde explosiones nucleares hasta la dinámica de fluidos— que eran analíticamente irresolubles.<sup>3</sup>

Hoy, sin embargo, confrontamos el Cuarto Paradigma: el descubrimiento científico intensivo en datos. En este nuevo estadio, la velocidad, el volumen y la variedad de los datos exceden la capacidad cognitiva humana para su procesamiento directo. El científico ya no mira a través del telescopio; el "telescopio" es un sensor digital que captura petabytes de datos, los cuales son almacenados en bases de datos distribuidas y minados por algoritmos.<sup>3,60</sup> El rol del investigador se desplaza de ser un observador directo a convertirse en un curador y analista de salidas computacionales.

Esta transición implica un cambio ontológico profundo. Grandes cantidades de datos experimentales fluyen desde fuentes masivas como el Gran Colisionador de Hadrones (LHC) en el CERN o secuenciadores genéticos de última generación, generando exabytes de información anualmente.<sup>1</sup> Extraer "insights" científicos de

estos datos es el desafío central, y para ello, los desarrollos más recientes en IA se vuelven esenciales. La metodología, por tanto, ya no trata solo de probar una hipótesis aislada, sino de gestionar el ciclo de vida completo de los datos, desde su captura y curación hasta su análisis y diseminación federada.<sup>5,61</sup>

## **1.2 El Debate sobre el "Fin de la Teoría"**

Un cisma metodológico crítico en la ciencia de datos moderna fue encendido por la provocadora afirmación de Chris Anderson en 2008, quien declaró que "el diluvio de datos hace obsoleto el método científico".<sup>6,62</sup> Anderson argumentó que, con suficientes datos, la correlación suplanta a la causalidad, y que la ciencia puede avanzar sin modelos coherentes o explicaciones mecanicistas unificadas. Esta perspectiva del "Fin de la Teoría" sugiere una ruptura radical con la investigación impulsada por hipótesis que ha definido la era post-Ilustración.

Bajo esta visión, en la era de la información de petabytes y la supercomputación, el método científico tradicional —basado en formular una hipótesis, diseñar un experimento y validarlo— se volvería anacrónico. En su lugar, lo que contaría serían los algoritmos sofisticados y las herramientas estadísticas capaces de cribar cantidades masivas de datos para encontrar patrones que pudieran transformarse en conocimiento.<sup>6</sup>

### **1.2.1 Crítica Epistemológica al Empirismo Ingenuo**

Sin embargo, un análisis riguroso revela que el "Big Data" no reemplaza la hipótesis, sino que cambia su génesis y ciclo de vida. Los enfoques puramente impulsados por los datos —a menudo descritos como "expediciones de pesca"— conllevan riesgos significativos de hallazgos espurios. En sistemas complejos con alta dimensionalidad, el ruido aleatorio puede disfrazarse fácilmente de señal, y las

correlaciones, por fuertes que sean, se vuelven exponencialmente más tenues al aumentar el tamaño de los datos si no hay una base causal.<sup>7</sup>

La crítica filosófica a la posición de Anderson enfatiza que la ciencia es fundamentalmente una actividad explicativa. Sin un modelo para explicar *por qué* las variables están correlacionadas, la predicción permanece frágil y no generalizable. Si la ciencia renuncia a la explicación en favor de la correlación ciega, corre el riesgo de degradarse a una actividad similar a la "colección de sellos", acumulando hechos sin estructura.<sup>8</sup>

### 1.3 Razonamiento Abductivo: La Síntesis Metodológica

La metodología propuesta rechaza el "Fin de la Teoría" como un empirismo ingenuo. En su lugar, aboga por un enfoque híbrido centrado en el razonamiento abductivo. A diferencia de la deducción (que garantiza conclusiones a partir de premisas verdaderas) o la inducción (que generaliza a partir de observaciones específicas), la abducción busca la "mejor explicación plausible" para un conjunto de datos incompleto (ver Tabla 3).<sup>9</sup>

**Tabla 3: Síntesis metodológica**

Tipo de Razonamiento	Descripción Metodológica	Aplicación en Ciencia de Datos e IA
Deductivo	Aplica principios generales a casos específicos para obtener conclusiones lógicamente ciertas.	Validación de modelos: Si la hipótesis es cierta, el modelo <i>debe</i> comportarse de manera X en el conjunto de prueba.

<b>Inductivo</b>	Identifica patrones a partir de observaciones específicas para formar generalizaciones probabilísticas.	Entrenamiento de modelos: El algoritmo "aprende" pesos y sesgos a partir de ejemplos de entrenamiento para generalizar a datos no vistos.
<b>Abductivo</b>	Determina la explicación más plausible basada en información incompleta.	Diagnóstico y generación de hipótesis: Ante un patrón anómalo en los datos, la IA sugiere la causa más probable (e.g., IA Co-científico). <sup>10</sup>

En el contexto de la investigación moderna, esto implica utilizar el aprendizaje automático para identificar patrones (fase inductiva) y luego construir hipótesis mecanicistas rigurosas para explicarlos (fase abductiva), las cuales son subsecuentemente probadas en conjuntos de datos independientes o mediante experimentos físicos (fase deductiva).<sup>6</sup>

## 2. Arquitecturas de Procesos y Ciclos de Vida en la Investigación

Para navegar la complejidad inherente a la investigación en IA, el campo ha adoptado y adaptado varios modelos de procesos provenientes de la ingeniería y la industria. Si bien estos modelos proporcionan un andamiaje para la gestión de proyectos, su aplicación en la investigación académica requiere modificaciones sustanciales para asegurar el rigor científico.

### 2.1 El Marco CRISP-DM: Estándar y Adaptación

El Proceso Estándar Inter-Industria para la Minería de Datos (CRISP-DM) permanece como la metodología más citada y utilizada para proyectos de ciencia de



datos (ver Tabla 4).<sup>12</sup> Desarrollado a finales de los años 90, descompone el ciclo de vida del proyecto en seis fases interconectadas e iterativas.

### 2.1.1 Fases del CRISP-DM y su Equivalente Científico

**Tabla 4: El Marco CRISP-DM**

Fase CRISP-DM	Tareas Industriales Típicas	Equivalente en Investigación Científica
<b>1. Comprensión del Negocio</b>	Definir objetivos, evaluar la situación, determinar criterios de éxito empresarial.	<b>Formulación del Problema:</b> Identificación de la brecha epistémica, definición de la pregunta de investigación y formulación de hipótesis. <sup>13</sup>
<b>2. Comprensión de los Datos</b>	Recolección inicial, descripción, exploración y verificación de calidad.	<b>Análisis Exploratorio de Datos (EDA):</b> Evaluación de la viabilidad de los datos para probar la hipótesis planteada.
<b>3. Preparación de los Datos</b>	Limpieza, construcción de atributos, integración y formateo ("80% del trabajo").	<b>Ingeniería de Características y Preprocesamiento:</b> Normalización rigurosa, manejo de valores atípicos y segmentación (train/val/test).
<b>4. Modelado</b>	Selección de técnicas, diseño de pruebas, construcción y evaluación del modelo.	<b>Diseño Experimental:</b> Entrenamiento de modelos, optimización de hiperparámetros y estudios de ablación.
<b>5. Evaluación</b>	Evaluación de resultados frente a objetivos de negocio, revisión del proceso.	<b>Validación y Análisis de Errores:</b> Pruebas de significancia estadística, análisis de fallos y revisión por

		pares de la metodología. <sup>13</sup>
<b>6. Despliegue</b>	Planificación del despliegue, monitoreo y reporte final.	<b>Diseminación y Publicación:</b> Redacción del paper, liberación de código reproducible y publicación de datasets (FAIR). <sup>13</sup>

A pesar de su utilidad, el CRISP-DM ha recibido críticas por ser demasiado rígido y orientado a "cascada" para la investigación moderna, que es altamente iterativa.<sup>14</sup> Además, carece de fases explícitas para la revisión ética y la auditoría de sesgos, componentes que ahora son obligatorios en la investigación rigurosa de IA.<sup>15</sup>

## 2.2 Alternativas Académicas: KDD y SEMMA

El proceso KDD (Knowledge Discovery in Databases) ofrece una alternativa más centrada en los datos y menos en el negocio. KDD enfatiza la transformación de datos crudos en "conocimiento" a través de una secuencia de selección, preprocesamiento, transformación, minería de datos e interpretación/evaluación.<sup>13</sup> Es a menudo preferido en la investigación de ciencias de la computación teórica donde la novedad reside en el método de extracción algorítmica más que en la aplicación práctica.

Por otro lado, SEMMA (Sample, Explore, Modify, Model, Assess), desarrollado por el instituto SAS, es un proceso secuencial enfocado en los pasos técnicos de la construcción del modelo.<sup>13</sup> Sin embargo, se considera demasiado estrecho para una metodología de investigación integral, ya que omite las fases críticas de contextualización (Comprensión del Negocio) y de impacto (Despliegue/Diseminación).

## 2.3 Hacia un Modelo Híbrido: CRISP-DS y Agile

La literatura reciente sugiere la necesidad de un Proceso Estándar Inter-Industria Generalizado para la Ciencia de Datos (GCRISP-DS), que permita interacciones dinámicas entre fases.<sup>18</sup> En la investigación científica, la fase de "Despliegue" es a menudo recursiva; un modelo publicado se convierte en la línea base para la siguiente iteración de investigación.

La metodología que se aboga en este texto integra principios Agile —sprints iterativos de prueba de hipótesis— dentro de las fases estructuradas de CRISP-DM. Esto permite a los investigadores "fallar rápido" cuando una hipótesis no está respaldada por la exploración temprana de datos, ahorrando recursos antes de comenzar un modelado profundo.<sup>14</sup> Este enfoque híbrido es esencial para gestionar la incertidumbre inherente a la investigación científica moderna.

### **3. La Crisis de Rigor y Reproducibilidad**

Una metodología es tan buena como la fiabilidad de sus resultados. La investigación en Inteligencia Artificial y Aprendizaje Automático (ML) se enfrenta actualmente a una severa "crisis de reproducibilidad", que corre paralela a la crisis de replicación en psicología y medicina.<sup>19</sup> Este capítulo detalla los fallos metodológicos que impulsan esta crisis y los estándares requeridos para mitigarlos.

#### **3.1 Anatomía del Fracaso: Fuga de Datos (Data Leakage)**

El error metodológico más pernicioso en la ciencia basada en ML es la fuga de datos (data leakage) —la introducción inadvertida de información del conjunto de prueba en el proceso de entrenamiento, lo que lleva a estimaciones de rendimiento "salvajemente optimistas".<sup>19</sup> Un estudio sistemático de la Universidad de Princeton identificó que este problema es sistémico y afecta a cientos de papers en docenas de

campos.<sup>19</sup>

Una metodología rigurosa debe controlar explícitamente los siguientes tipos de fuga:

1. **Fuga de Preprocesamiento:** Ocurre cuando se realizan transformaciones como la normalización, la selección de características o la imputación de valores perdidos sobre el conjunto de datos *completo* antes de dividirlo en entrenamiento y prueba. Esto contamina los datos de entrenamiento con estadísticas globales que incluyen información del conjunto de prueba.<sup>19</sup>
  - *Solución Metodológica:* Dividir los datos antes de cualquier toque. Calcular estadísticas solo en el conjunto de entrenamiento y aplicarlas al de prueba.
2. **Fuga Temporal:** En datos de series temporales, entrenar con datos futuros para predecir eventos pasados. El uso de validación cruzada aleatoria (k-fold) en series temporales destruye la causalidad temporal.<sup>19</sup>
  - *Solución Metodológica:* Uso estricto de validación cruzada de "ventana rodante" o división temporal (entrenar en pasado, probar en futuro).
3. **Fuga de Características Ilegítimas (Proxies):** Incluir características que son proxies casi perfectos de la variable objetivo (por ejemplo, incluir "duración de la cirugía" en un modelo que predice "complicaciones quirúrgicas", cuando las complicaciones inherentemente extienden la duración).<sup>19</sup>

### 3.2 Listas de Verificación de Reproducibilidad (Checklists)

Para combatir estos errores y estandarizar el rigor, conferencias de primer nivel como NeurIPS e ICML han instituido Listas de Verificación de Reproducibilidad. La adopción de estas listas no es un mero trámite burocrático, sino una exigencia metodológica fundamental. Los componentes clave incluyen:

- **Disponibilidad del Código:** El envío de código limpio y ejecutable con dependencias especificadas es obligatorio. Un paper sin código es, en muchos casos, no verificable.<sup>21</sup>
- **Transparencia de Hiperparámetros:** Se debe reportar el rango detallado de hiperparámetros buscados y el método de selección (e.g., búsqueda en rejilla vs. optimización bayesiana).<sup>21</sup>
- **Control de la Aleatoriedad:** Especificar las semillas aleatorias (random seeds) y realizar múltiples ejecuciones para reportar la media y la varianza, en lugar de una sola "ejecución heroica" que podría ser una aberración estadística.<sup>21</sup>
- **Infraestructura Computacional:** Reportar el hardware utilizado (e.g., tipo de GPU), ya que las diferencias en hardware pueden llevar a variaciones sutiles en los cálculos de punto flotante.<sup>21</sup>

Metodológicamente, esto mueve al campo lejos de los "papers anuncio" —que reclaman resultados de Estado del Arte (SOTA) basados en experimentos oscuros— hacia "papers científicos" donde el proceso es auditable y refutable.

## 4. Documentación como Metodología Científica

En la ciencia tradicional, la sección de "Materiales y Métodos" de un artículo era suficiente documentación. En la IA, donde el "material" es un conjunto de datos dinámico y el "método" es un algoritmo opaco, la documentación debe ser mucho más granular y estructurada. Los conceptos de Datasheets (Hojas de Datos) y Model Cards (Tarjetas de Modelo) han emergido como artefactos metodológicos esenciales.

### 4.1 Datasheets for Datasets (Hojas de Datos para Conjuntos de Datos)

Propuesto por Gebru et al., este marco estandariza la documentación del ciclo de vida de los datos.<sup>22</sup> Una metodología de investigación que ignora la procedencia de los datos es fundamentalmente defectuosa. Un Datasheet riguroso debe responder:

- **Motivación:** ¿Por qué se creó el conjunto de datos? ¿Quién lo financió?
- **Composición:** ¿Qué representan las instancias? ¿Existen subpoblaciones? ¿Es inclusivo el conjunto de datos o presenta sesgos demográficos?
- **Proceso de Colección:** ¿Cómo se adquirieron los datos? ¿Estuvieron involucrados sujetos humanos? ¿Se obtuvo consentimiento informado?
- **Preprocesamiento:** ¿Qué limpieza, muestreo o anonimización se realizó?

Para el investigador de IA, el Datasheet es una herramienta de **auditabilidad**. Previene el mal uso de conjuntos de datos en contextos para los cuales no fueron diseñados (por ejemplo, usar un dataset de rostros occidentales para entrenar un sistema de reconocimiento facial global).<sup>24</sup>

## 4.2 Model Cards for Model Reporting (Tarjetas de Modelo)

Complementando a los Datasheets, las Model Cards (Mitchell et al.) proporcionan un reporte estandarizado sobre el modelo entrenado.<sup>25</sup> Una Model Card va más allá de las métricas de precisión global para documentar:

- **Uso Previsto:** ¿Para qué fue diseñado el modelo? ¿Cuáles son los usos "fuera de alcance" (out-of-scope)?
- **Factores:** ¿Cómo varía el rendimiento del modelo a través de diferentes grupos demográficos, condiciones ambientales o atributos técnicos?
- **Métricas:** ¿Qué métricas de rendimiento se eligieron y por qué? (e.g., priorizar la sensibilidad sobre la precisión en diagnósticos médicos).
- **Consideraciones Éticas:** ¿Existen riesgos de sesgo, vigilancia o doble uso?

La metodología de "Model Reporting" obliga al investigador a confrontar las limitaciones de su trabajo. Desplaza el objetivo de maximizar una métrica única (como la exactitud) hacia la caracterización del *comportamiento* del sistema en un espacio multidimensional.<sup>27</sup>

## 5. La IA Co-Científica y Métodos Generativos

La metodología de la investigación en IA está experimentando una evolución recursiva: la IA se está utilizando para mejorar la investigación en IA. Los modelos generativos y los Grandes Modelos de Lenguaje (LLMs) están alterando las fases de generación de hipótesis y revisión de literatura.

### 5.1 Generación Automatizada de Hipótesis

Sistemas como el AI Co-scientist de Google, construido sobre Gemini 2.0, demuestran el potencial de los sistemas multi-agente para funcionar como colaboradores científicos virtuales.<sup>11</sup> Estos sistemas utilizan un ciclo de "método científico" interno:

1. **Generación:** Un agente propone una hipótesis novedosa basada en la literatura existente.
2. **Reflexión:** Un agente crítico evalúa la hipótesis por su plausibilidad y novedad.
3. **Ranking/Evolución:** Las hipótesis se clasifican mediante torneos simulados (usando clasificaciones Elo) y se evolucionan para mejorar su calidad.<sup>11</sup>
4. **Proximidad:** Asegurar que la hipótesis esté fundamentada en trabajo previo válido.

Esto introduce una nueva capacidad metodológica: cribado de hipótesis de alto rendimiento. Un investigador puede generar cientos de direcciones de

investigación potenciales y usar la IA para filtrar las más prometedoras antes de comprometer recursos experimentales.

## 5.2 Riesgos de los LLMs en la Investigación

Sin embargo, la integración de LLMs introduce riesgos de validez significativos que la metodología debe gestionar:

- **Alucinación:** Los LLMs pueden generar citas plausibles pero inexistentes o datos falsos. Confiar en un LLM para la revisión de literatura sin verificación manual es un error metodológico grave.<sup>28</sup>
- **Erosión de Habilidades Críticas:** La dependencia excesiva de la IA para escribir e idear puede atrofiar la capacidad del investigador para realizar síntesis profundas y análisis crítico.<sup>29</sup>
- **Contaminación de la Revisión por Pares:** El uso de LLMs para escribir revisiones por pares socava la integridad del proceso de evaluación. Estudios indican un aumento significativo en contenido generado por IA en revisiones de conferencias, planteando preocupaciones sobre la confidencialidad y la calidad de la crítica.<sup>31</sup>

Por lo tanto, el enfoque general restringe la IA Generativa al rol de un *asistente* —una herramienta para lluvia de ideas, resumen y codificación— mientras reserva el rol de *autor* y *juez* para el científico humano.<sup>30</sup>

## 6. Ética, Gobernanza y el Nuevo Marco Regulatorio

La metodología en la investigación de IA ya no puede separarse de la ética. El potencial de los sistemas de IA para causar daño a escala —a través de sesgos, desinformación o toma de decisiones autónoma— ha necesitado la integración de



marcos de gobernanza directamente en el ciclo de vida de la investigación.

## 6.1 La Declaración de Impacto Amplio (Broader Impact Statement)

Desde 2020, NeurIPS ha requerido que los autores incluyan una Declaración de Impacto Amplio, reflexionando sobre las consecuencias sociales potenciales, tanto positivas como negativas, de su trabajo.<sup>34</sup> Este requisito obliga a los investigadores a participar en una gobernanza anticipatoria. No es suficiente construir un algoritmo de reconocimiento facial; el investigador debe analizar cómo podría ser mal utilizado para la vigilancia o cómo podría fallar en poblaciones subrepresentadas.<sup>36</sup>

Aunque algunos ven esto como una barrera burocrática, es metodológicamente esencial para determinar la *validez* de una contribución. Un modelo que logra una alta precisión pero se basa en mano de obra explotada o viola normas de privacidad es científica y éticamente defectuoso.<sup>37</sup>

## 6.2 Marcos Regulatorios como Restricciones Metodológicas

Las metodologías de investigación deben alinearse ahora con marcos legales emergentes:

- **Ley de IA de la UE (EU AI Act):** Esta legislación categoriza los sistemas de IA por nivel de riesgo. Los sistemas de "Alto Riesgo" (e.g., dispositivos médicos, biometría) enfrentan requisitos estrictos de gobernanza de datos, supervisión humana y precisión.<sup>38</sup> Los investigadores que desarrollan tales sistemas deben documentar su cumplimiento con estos requisitos *durante* la fase de investigación, no después. La Ley prohíbe ciertas direcciones de investigación por completo, como la manipulación subliminal o la puntuación social (social

scoring).<sup>39</sup>

- **Marco de Gestión de Riesgos de IA del NIST (AI RMF):** El RMF del NIST proporciona una estructura voluntaria pero influyente para gestionar riesgos de IA. Sus funciones centrales —Gobernar, Mapear, Medir, Gestionar— proporcionan una plantilla para las operaciones de investigación.<sup>40</sup> La función de "Medir", por ejemplo, estandariza cómo cuantificamos la "confiabilidad", la "robustez" y el "sesgo".

## 6.3 Ciclos de Vida de IA Responsable

Líderes de la industria como Microsoft y Google han formalizado ciclos de vida de IA Responsable. Estos marcos enfatizan principios como "Justicia", "Confiabilidad y Seguridad", "Privacidad", "Inclusión", "Transparencia" y "Rendición de Cuentas".<sup>42</sup>

Adoptar estos principios significa que un proyecto de investigación no está completo cuando se maximiza la métrica de precisión. Está completo cuando el modelo ha sido probado bajo estrés para detectar sesgos, se han evaluado sus vulnerabilidades de seguridad (e.g., robustez adversarial) y se ha documentado su proceso de toma de decisiones.<sup>43</sup>

Herbert Simon, uno de los padres fundadores de la IA, imaginó una "Ciencia de lo Artificial" —una disciplina que estudia los objetos creados por el diseño humano con el mismo rigor con que las ciencias naturales estudian el mundo físico. Hoy, esa visión es más relevante que nunca. A medida que los sistemas de IA se vuelven más complejos, exhiben comportamientos emergentes que no pueden predecirse solo a partir de su código; deben estudiarse empíricamente.<sup>30</sup>

Al adherirse a estándares rigurosos de reproducibilidad, documentación y previsión ética, aseguramos que el "Cuarto Paradigma" produzca no solo datos, sino

sabiduría duradera. El enfoque detallado aquí no es simplemente un conjunto de reglas; es un compromiso con la integridad de la empresa científica en una era algorítmica.

# **Capítulo 3.**

## **Prolegómenos: La Crisis Epistemológica del Cuarto Paradigma**

### **1. Rigor metodológico de la inteligencia artificial**

La ciencia, en su concepción más noble, es la búsqueda sistemática de la verdad a través de la observación, la hipótesis y la experimentación. Sin embargo, en los albores del siglo XXI, la naturaleza misma de este proceso ha sufrido una transformación radical, una metamorfosis impulsada por la disponibilidad de datos masivos y la capacidad de cómputo ubicua. Nos encontramos ante lo que Jim Gray, el visionario científico de la computación y ganador del Premio Turing, denominó el Cuarto Paradigma de la exploración científica: la ciencia intensiva en datos (eScience).<sup>1</sup>

La inteligencia artificial (IA) y la ciencia de datos han dejado de ser subdisciplinas oscuras de las ciencias de la computación para convertirse en los lentes a través de los cuales observamos y manipulamos la realidad. Desde la predicción del plegamiento de proteínas en biología molecular hasta la modelización de riesgos climáticos, la IA actúa como un "motor de difusión" que acelera el descubrimiento a través de las fronteras disciplinarias.<sup>3</sup> No obstante, esta aceleración ha tenido un costo: la proliferación de una "alquimia digital" donde la correlación se confunde con la causalidad, y donde la optimización de métricas en conjuntos de datos estáticos sustituye a la generación de conocimiento generalizable.

La premisa central de esta obra es que la IA no es simplemente una ingeniería de software aplicada, ni es una caja negra mágica que exime al investigador de la

responsabilidad intelectual. Por el contrario, la investigación en IA exige un rigor metodológico *mayor* que la ciencia tradicional, precisamente porque sus herramientas son estocásticas, opacas y propensas a amplificar sesgos humanos de formas sutiles e insidiosas.<sup>4</sup> Mientras que Newton podía explicar la gravedad con la elegante simplicidad de  $F = G \frac{m_1 m_2}{r^2}$ , un investigador moderno en aprendizaje profundo se enfrenta a modelos con miles de millones de parámetros cuya lógica interna es a menudo inescrutable para sus propios creadores.<sup>4</sup>

## 1.1. La Evolución de la Indagación Científica

Para comprender la posición actual de la IA, debemos situarla en el arco histórico del desarrollo científico, tal como lo conceptualizó Gray y fue expandido por sus colegas de Microsoft Research y la comunidad académica global (ver Tabla 5).<sup>1</sup>

**Tabla 5: Paradigmas de Jim Gray**

Paradigma	Descripción	Naturaleza de la Investigación	Herramientas Principales
<b>Primer Paradigma</b>	Ciencia Empírica	Descriptiva: Observación de fenómenos naturales (ej. Brahe observando estrellas).	Ojos, telescopios rudimentarios, bitácoras.
<b>Segundo Paradigma</b>	Ciencia Teórica	Analítica: Generalización a través de leyes y modelos matemáticos (ej. Newton, Maxwell).	Cálculo, ecuaciones diferenciales, teoría.

<b>Tercer Paradigma</b>	Ciencia Computacional	Simulación: Modelado de sistemas complejos demasiado difíciles para el análisis puro (ej. dinámica de fluidos).	Supercomputadoras, métodos numéricos, simulaciones de Monte Carlo.
<b>Cuarto Paradigma</b>	<b>Ciencia Intensiva en Datos</b>	<b>Exploratoria/Predictiva:</b> Unificación de teoría, experimento y simulación a través del análisis de datos masivos.	Bases de datos distribuidas, <b>Machine Learning</b> , <b>IA</b> , Estadística Bayesiana.

La transición al Cuarto Paradigma implica que los científicos ya no miran directamente a través de sus instrumentos; miran a través de bases de datos. Los datos son capturados por instrumentos o generados por simuladores, procesados por software y almacenados en infraestructuras federadas.<sup>2</sup> En este contexto, la IA no es el objeto de estudio final, sino el *instrumento* de descubrimiento. El problema surge cuando aplicamos metodologías del tercer paradigma (simulación controlada) o prácticas de ingeniería de software a problemas del cuarto paradigma, que son inherentemente ruidosos, sesgados y no estacionarios.<sup>3</sup>

## 1.2. La Falsa Equivalencia: Ingeniería vs. Investigación

Uno de los obstáculos más perniciosos para el avance científico en este campo es la confusión categórica entre la Ingeniería de Software y la Investigación en Ciencia de Datos/IA. Esta distinción no es semántica; es metodológica y define el éxito o fracaso de un proyecto.

La ingeniería de software se centra en la construcción de sistemas fiables,

mantenibles y escalables. Su ciclo de vida es lineal o ágil, enfocado en la entrega continua de funcionalidad.<sup>8</sup> Por el contrario, la ciencia de datos es intrínsecamente exploratoria y experimental. Su objetivo no es construir un sistema que "funcione" en el sentido de que no se cuelgue, sino descubrir una verdad oculta en los datos o construir un modelo que generalice a situaciones no vistas.<sup>8</sup>

El gerente de ingeniería que exige "sprints" de dos semanas para un descubrimiento científico está cometiendo un error epistemológico. Como señalan diversos expertos en la industria, intentar forzar el ciclo de vida de la IA (iterativo, incierto, basado en hipótesis) dentro de los marcos rígidos de la ingeniería de software tradicional conduce a una deuda técnica masiva y a modelos defectuosos (ver Tabla 6).<sup>9</sup>

**Tabla 6: Divergencias Metodológicas Fundamentales <sup>8</sup>**

Dimensión	Ingeniería de Software Tradicional	Investigación en IA y Ciencia de Datos
Objetivo Primario	Funcionalidad y Fiabilidad del Sistema.	Extracción de Conocimiento y Generalización Predictiva.
Enfoque de Calidad	Minimización de bugs, cobertura de código.	Minimización de sesgo/varianza, robustez estadística.
Incertidumbre	Baja: El comportamiento es determinista.	Alta: El comportamiento es estocástico y probabilístico.
Gestión de Datos	Datos transaccionales, esquemas fijos.	Datos no estructurados, ruidosos, requieren limpieza masiva.

<b>Rol Humano</b>	Desarrolladores, QA, Product Managers.	Investigadores, <b>Expertos en el Dominio (SME)</b> , Ingenieros de Datos.
<b>Ciclo de Vida</b>	Definición -> Diseño -> Desarrollo -> Test.	Hipótesis -> Experimento -> Análisis -> <b>Retorno al Diseño.</b>

La investigación en IA requiere la reintroducción del Método Científico clásico: observación, formulación de hipótesis, predicción, experimentación y análisis.<sup>10</sup> No basta con lanzar datos a una red neuronal y esperar que "aprenda"; se debe tener una "corazonada creíble" basada en la teoría previa y diseñar experimentos para falsear esa corazonada.<sup>10</sup>

## 2. La Anatomía de la Crisis de Reproducibilidad

Si el advenimiento del Cuarto Paradigma representa la promesa de la IA, la "Crisis de Reproducibilidad" representa su mayor amenaza existencial. La comunidad científica ha comenzado a reconocer que una proporción alarmante de los resultados publicados en conferencias de prestigio (NeurIPS, ICML, AAAI) no pueden ser replicados por investigadores independientes.<sup>13</sup>

Esta crisis no es exclusiva de la computación —la psicología y la biomedicina han sufrido sus propios "inviernos de replicación"— pero en la IA adquiere matices únicos debido a la naturaleza del software y los datos.<sup>15</sup> La reproducibilidad se define estrictamente como la capacidad de un investigador para duplicar los resultados de un estudio previo utilizando los mismos materiales (código y datos).<sup>17</sup> La replicabilidad, un estándar aún más alto, implica llegar a las mismas conclusiones recolectando nuevos datos y realizando nuevos experimentos.<sup>18</sup>



## 2.1. Los Pecados Capitaes de la Investigación en IA

A través de la revisión de la literatura y el análisis de retractaciones recientes, hemos identificado una taxonomía de fallos metodológicos que busca erradicar:

1. **P-Hacking y Dragado de Datos:** Al igual que en las ciencias sociales, los investigadores en IA a menudo ejecutan cientos de experimentos variando hiperparámetros y arquitecturas, pero solo reportan la configuración que obtuvo el mejor resultado en el conjunto de prueba. Esto viola los principios básicos de la estadística inferencial, inflando la tasa de falsos positivos y presentando ruido aleatorio como un descubrimiento significativo.<sup>13</sup> El valor  $p$  pierde su significado si no se corrige por comparaciones múltiples.
2. **HARKing (Hypothesizing After the Results are Known):** Presentar una hipótesis post-hoc como si hubiera sido la motivación original del estudio. Esto invierte el método científico, convirtiendo la investigación confirmatoria en exploratoria sin las advertencias necesarias.<sup>13</sup>
3. **La "Semilla de la Suerte" (Lucky Seed):** En el aprendizaje profundo y el aprendizaje por refuerzo, la inicialización aleatoria de los pesos de la red puede tener un impacto drástico en el rendimiento. Un estudio puede declarar que el Algoritmo A es superior al Algoritmo B simplemente porque tuvo suerte en la inicialización, no por una superioridad algorítmica intrínseca. Sin realizar múltiples ejecuciones con diferentes semillas y reportar la varianza, los resultados son anecdóticos, no científicos.<sup>18</sup>
4. **Fuga de Datos (Data Leakage):** Quizás el error más común y devastador. Ocurre cuando información del futuro (del conjunto de prueba) se filtra insidiosamente en el proceso de entrenamiento. Ejemplos documentados incluyen la normalización de datos antes de la separación train/test, o el uso de

características que son proxies directos de la variable objetivo (como incluir la duración de la estancia hospitalaria para predecir la gravedad de una enfermedad).<sup>19</sup>

## **Estudio de Caso: El Fracaso de la Autocorrección en LLMs**

Un ejemplo contemporáneo de fallo metodológico es la evaluación de la capacidad de "autocorrección" en Grandes Modelos de Lenguaje (LLMs). Investigaciones recientes han demostrado que los LLMs sufren de una "falla de autocorrección intrínseca" cuando intentan validar sus propias respuestas dentro del mismo contexto conversacional, confirmando sus errores más del 90% de las veces debido al sesgo de confirmación. Sin embargo, cuando se evalúa la misma respuesta en un entorno de "sala limpia" (clean room) —un contexto nuevo sin el historial previo— la capacidad de detección de errores mejora dramáticamente.<sup>21</sup> Los estudios que no controlan esta variable de contexto están, esencialmente, midiendo el sesgo del modelo en lugar de su capacidad de razonamiento, un error clásico de validez de constructo.

## **2.2. La Respuesta Institucional: Listas de Verificación y Retracciones**

Para combatir esta entropía metodológica, conferencias como NeurIPS introdujeron en 2019 programas de reproducibilidad obligatorios, incluyendo listas de verificación (checklists) que exigen a los autores declarar la disponibilidad de código, la especificación de hiperparámetros y los detalles de los recursos de cómputo.<sup>14</sup>

El análisis de la eficacia de estas medidas ofrece una visión mixta. Si bien el

cumplimiento autodeclarado es alto (alrededor del 62-74% de los artículos afirman cumplir con los requisitos y proporcionar código), la realidad empírica es menos alentadora.<sup>22</sup> Un estudio que intentó implementar manualmente 255 artículos encontró que la disponibilidad del código no garantiza la reproducibilidad si la descripción algorítmica en el papel difiere de la implementación, o si los datos no son accesibles debido a restricciones de propiedad o privacidad.<sup>23</sup>

Más alarmante aún es el aumento exponencial de las retractaciones. Según datos de *Retraction Watch*, la tasa de retractaciones ha pasado de 1 en 5,000 artículos en 2002 a 1 en 500 en 2023.<sup>24</sup> El auge de las "fábricas de artículos" (paper mills) y la presión por publicar han creado un entorno donde el fraude y el error honesto proliferan. Incluso herramientas de IA diseñadas para detectar artículos retractados han demostrado ser poco fiables, alucinando referencias o fallando en identificar trabajos desacreditados.<sup>25</sup> Esto subraya la necesidad de una vigilancia humana experta y una metodología que priorice la calidad sobre la cantidad.

### **3. El Marco Metodológico Propuesto: El Modelo Anidado y el Diseño Experimental**

Adoptamos y expandimos el Modelo Anidado (Nested Model) para el diseño y validación de sistemas de IA, una estructura que obliga al investigador a validar sus supuestos en cuatro niveles jerárquicos antes de reclamar el éxito.<sup>26</sup>

#### **3.1. Las Cuatro Capas de Validación**

El error fundamental en muchos proyectos de ciencia de datos es saltar directamente a la optimización de algoritmos sin validar el problema. El Modelo

Anidado estructura la investigación de la siguiente manera:

1. **Capa de Dominio (Domain Layer):** ¿Estamos resolviendo el problema correcto? Aquí, la metodología requiere la interacción con Expertos en la Materia (SMEs). Un algoritmo que optimiza el flujo de tráfico es inútil si no considera las regulaciones urbanas o la psicología del conductor. La validación aquí es cualitativa y conceptual.<sup>11</sup>
2. **Capa de Datos (Data Layer):** ¿Son los datos una representación fiel del dominio? Aquí abordamos la calidad, el sesgo y la procedencia de los datos. Se deben aplicar técnicas de "hojas de datos para conjuntos de datos" (Datasheets for Datasets) para documentar la creación, composición y limitaciones legales/éticas de los datos.<sup>27</sup>
3. **Capa de Modelo (Model Layer):** ¿Es el método capaz de capturar la estructura de los datos? Aquí es donde entra la selección de algoritmos (CNN, Transformer, Random Forest). La validación implica verificar que el modelo puede sobreajustarse (overfit) a un subconjunto pequeño de datos, demostrando que tiene la capacidad de aprendizaje necesaria.<sup>29</sup>
4. **Capa de Predicción (Prediction Layer):** ¿El modelo generaliza a datos no vistos? Esta es la prueba de fuego estadística. Implica el uso riguroso de conjuntos de retención (hold-out sets), validación cruzada anidada y pruebas de estrés fuera de distribución (OOD).<sup>19</sup>

### 3.2. Diseño de Experimentos (DoE) en el Silicio

La investigación en IA debe recuperar la disciplina del Diseño de Experimentos (DoE) utilizada en la ingeniería industrial y la agronomía. En lugar de la prueba y error aleatoria (o la "gradiente descendente de estudiante de posgrado"), el DoE permite explorar sistemáticamente el espacio de hiperparámetros y

configuraciones arquitectónicas.<sup>30</sup>

El uso de técnicas como la Optimización Bayesiana o el muestreo de hipercubo latino permite a los investigadores modelar la superficie de respuesta del rendimiento del algoritmo de manera eficiente. Además, el DoE obliga a considerar las interacciones entre variables (ej. ¿cómo interactúa la tasa de aprendizaje con el tamaño del lote?), algo que la optimización "uno a la vez" (one-at-a-time) ignora (ver Tabla 7).<sup>30</sup>

**Tabla 7: Estrategias de Validación Experimental <sup>19</sup>**

Estrategia	Descripción	Propósito Metodológico
<b>Validación Cruzada K-Fold</b>	Dividir datos en K partes, entrenar en K-1, probar en 1.	Estimar la varianza del rendimiento del modelo y maximizar uso de datos.
<b>Conjunto de Retención (Hold-out)</b>	Separar una porción de datos al inicio y <i>nunca</i> tocarla hasta el final.	Simular el despliegue en el mundo real y evitar el sobreajuste al conjunto de validación.
<b>Pruebas de Estrés (Adversarial)</b>	Introducir ruido o perturbaciones en los datos de entrada.	Evaluar la robustez y seguridad del modelo ante entradas maliciosas o corruptas.
<b>Ablación (Ablation Studies)</b>	Eliminar componentes del modelo secuencialmente.	Establecer causalidad interna: ¿qué parte del modelo contribuye al rendimiento?
<b>Evaluación "Clean Room"</b>	Evaluar respuestas sin contexto previo (específico para LLMs).	Evitar sesgos de confirmación y medir la capacidad real de

		razonamiento. <sup>21</sup>
--	--	-----------------------------

## 4. Estructura de un Plan de Estudios de Posgrado en Investigación con IA (ejes temáticos)

La secuencia del plan de estudios refleja el flujo lógico de un proyecto de investigación riguroso, integrando los syllabus de instituciones líderes como Carnegie Mellon, Stanford y Georgia Tech.<sup>33</sup>

### Semestre I: Fundamentos y Filosofía

- **1: La Ciencia en la Era del Algoritmo.** Historia del Cuarto Paradigma, la distinción ciencia/ingeniería.
- **2: Epistemología de la Inteligencia Artificial.** ¿Qué significa que una máquina "sepa" algo? Inducción vs. Deducción en ML. Causalidad vs. Correlación (Inferencia Causal).<sup>36</sup>

### Semestre II: Diseño y Preparación

- **3: Definición del Problema y Revisión de Literatura.** Cómo pasar de una idea vaga a una pregunta de investigación falsable. El uso de herramientas de IA para la revisión bibliográfica y sus peligros (alucinaciones).<sup>37</sup>
- **4: Ética y Regulación como Diseño.** La ética no es un anexo. Privacidad diferencial, equidad algorítmica (Fairness), y cumplimiento con regulaciones (EU AI Act) desde la fase de diseño.<sup>26</sup>
- **5: Metodología de los Datos.** Recolección, limpieza, anotación y la gestión del ciclo de vida de los datos (Data Lifecycle). Detección de sesgos en datasets.<sup>27</sup>

## Semestre III: Experimentación y Modelado (El Núcleo Técnico)

- **6: Diseño de Experimentos Computacionales.** DoE, selección de métricas (más allá del Accuracy), y la definición de líneas base (Baselines) honestas.<sup>17</sup>
- **7: Algoritmos y Comparación.** Metodologías específicas para Supervisado, No Supervisado y Refuerzo. Pruebas de significancia estadística para comparar modelos (Test de McNemar, Wilcoxon, etc.).<sup>15</sup>
- **8: Reproducibilidad y Gestión de Flujos de Trabajo.** Herramientas (Docker, MLflow), control de versiones de datos (DVC) y listas de verificación de conferencias.<sup>14</sup>

## Semestre IV: Análisis y Comunicación

- **9: Interpretabilidad y Explicabilidad (XAI).** Métodos post-hoc (SHAP, LIME) vs. modelos intrínsecamente interpretables. Validación de explicaciones con humanos.<sup>4</sup>
- **10: Escritura Científica y Publicación.** Estructura de un paper de IA. Cómo reportar resultados negativos. El proceso de revisión por pares y cómo responder a los revisores (Rebuttal).<sup>42</sup>
- **11: La Tesis Doctoral en IA.** Guía específica para estudiantes de doctorado. Cómo elegir un asesor, cómo seleccionar un tema con "ventaja competitiva", y la gestión de la salud mental en un campo hiper-competitivo.<sup>38</sup>

# 5. El Factor Humano: Consejos para el Investigador Novel

## 5.1. La Elección del Problema y la Ventaja Competitiva

Un error común entre los estudiantes noveles es intentar competir directamente con los laboratorios industriales (Google DeepMind, OpenAI) en el entrenamiento de modelos masivos. Esto es una batalla perdida debido a la disparidad de recursos computacionales. El consejo de expertos como Andrej Karpathy y académicos de primer nivel es claro: busque problemas donde tenga una ventaja injusta.<sup>44</sup> Esta ventaja puede ser un acceso único a un conjunto de datos (ej. datos clínicos de un hospital asociado), una colaboración interdisciplinaria profunda (ej. trabajar con biólogos o historiadores), o un enfoque teórico novedoso que no requiera miles de GPUs. La simplicidad a menudo supera a la complejidad; las ideas simples que resisten el paso del tiempo son más valiosas que las arquitecturas barrocas que solo funcionan en un conjunto de datos específico.<sup>44</sup>

## **5.2. La Responsabilidad Ética y el Uso de IA Generativa**

El estudiante moderno cuenta con herramientas poderosas como los asistentes de codificación y redacción basados en LLMs. Las universidades están empezando a emitir directrices claras: el uso de IA es bienvenido para la lluvia de ideas o la corrección de código, pero el autor humano es el único responsable de la integridad del trabajo.<sup>38</sup> Un investigador no puede culpar a la IA por una cita alucinada o un error en la lógica matemática. La "responsabilidad algorítmica" comienza con la responsabilidad personal del científico. El uso de estas herramientas debe ser transparente y declarado, fomentando un pensamiento crítico sobre lo que la máquina genera.<sup>47</sup>



## Capítulo 4.

# La epistemología de lo artificial: un marco integral para la metodología de la investigación científica en inteligencia artificial y ciencia de datos

### 1. La convergencia entre computación y empirismo

La aparición de la Inteligencia Artificial (IA) y la Ciencia de Datos ha precipitado un cambio fundamental en la ontología del descubrimiento científico, representando lo que Jim Gray denominó famosamente el "Cuarto Paradigma" de la ciencia.<sup>1</sup> Históricamente, el método científico ha evolucionado a través de tres fases distintas: la descripción empírica de fenómenos naturales, la formulación teórica de leyes (por ejemplo, las leyes de Newton, las ecuaciones de Maxwell) y la simulación computacional de sistemas complejos.<sup>2</sup> La era actual, caracterizada por un descubrimiento científico intensivo en datos, sintetiza estos predecesores pero presenta un desafío epistemológico distinto: la conciliación de modelos predictivos de alta dimensión, a menudo opacos, con las exigencias rigurosas y explicativas de la investigación científica tradicional.

Este informe articula una metodología integral para aplicar el método científico a la IA y la Ciencia de Datos. Va más allá del enfoque centrado en la ingeniería de maximizar métricas predictivas (por ejemplo, precisión, puntuación F1) para establecer un marco riguroso para la generación de hipótesis, el diseño experimental, la validación estadística y la explicación causal. La tesis central planteada aquí es que, para que la IA sirva como un instrumento genuino de la ciencia —y no simplemente una herramienta de automatización— debe estar sometida a estándares de control y protocolos de reproducibilidad que definen las ciencias naturales.

#### 1.1 La tensión entre predicción y explicación

Una dicotomía crítica en la ciencia de datos moderna, articulada por Leo Breiman como las "Dos Culturas" y refinada aún más por Galit Shmueli, es la distinción entre *explicación* y *predicción*.<sup>4</sup> La modelización explicativa, dominante en campos como la econometría y la bioestadística, busca poner a prueba hipótesis causales y estimar parámetros que reflejan los constructos teóricos subyacentes. En cambio, el modelado predictivo —el bastión del aprendizaje automático (ML) moderno— prioriza la minimización del error fuera de muestra, a menudo a costa de la interpretabilidad y la fidelidad teórica.<sup>4</sup>

Esta divergencia tiene profundas implicaciones metodológicas. En un marco explicativo, un modelo se valida por su bondad de ajuste y la significación estadística de sus coeficientes. En un marco predictivo, la validez se determina por el rendimiento de generalización sobre datos ocultos.<sup>7</sup> Sin embargo, a medida que los sistemas de IA se despliegan cada vez más en dominios científicos de alto riesgo—desde el descubrimiento de fármacos<sup>8</sup> hasta la modelización climática—el enfoque predictivo de la "caja negra" se vuelve insuficiente. Por tanto, una metodología científica para la IA debe salvar esta brecha, integrando el poder inductivo del aprendizaje automático con el rigor deductivo de las pruebas de hipótesis.

## 1.2 El estado epistémico de la investigación basada en datos

La provocación del "fin de la teoría", que sugería que la cantidad suficiente de datos hace obsoleto el método científico, ha sido en gran medida rechazada por la comunidad de la filosofía de la ciencia.<sup>10</sup> Big Data no reemplaza el método científico; más bien, altera el mecanismo de generación de hipótesis. La ciencia basada en datos funciona como un motor de detección de anomalías, identificando correlaciones que requieren explicación teórica.<sup>12</sup> En este contexto, los modelos de aprendizaje automático actúan como "teorías blandas"—espacios complejos de hipótesis restringidos por sesgos inductivos.<sup>13</sup>

La metodología detallada en este informe operacionaliza esta perspectiva. Trata el entrenamiento de un modelo de aprendizaje automático no como la conclusión de la investigación, sino como un experimento en sí mismo: una sonda en la variedad de alta dimensión de los datos. Las secciones siguientes definen cómo estructurar esta investigación, asegurando que el "Científico de IA" cumpla con los

estándares de reproducibilidad, control y falsabilidad.

## 2. Modelos de procesos estructurados: desde los estándares industriales hasta el rigor científico

La práctica de la ciencia de datos ha estado guiada en gran medida por modelos de procesos derivados de aplicaciones industriales. Aunque estos marcos proporcionan la estructura necesaria, a menudo carecen de los mecanismos específicos necesarios para la validez científica.

### 2.1 Crítica de los modelos industriales estándar

El Proceso Estándar Interindustrial para Minería de Datos (CRISP-DM) y el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) son las metodologías dominantes en el campo.<sup>15</sup>

#### 2.1.1 CRISP-DM

CRISP-DM describe un ciclo de vida cíclico compuesto por seis fases: Comprensión del Negocio, Comprensión de Datos, Preparación de Datos, Modelado, Evaluación y Despliegue.<sup>15</sup>

- **Fortalezas:** Enfatiza la naturaleza iterativa de los proyectos de datos y la importancia del contexto del dominio ("Comprensión Empresarial").<sup>18</sup>
- **Limitaciones científicas:** CRISP-DM es teleológico; su objetivo es *la utilidad* (resolver un problema empresarial) más que *la verdad* (verificar una hipótesis). La fase de "Evaluación" se centra en si el modelo cumple con los criterios de éxito empresarial, no en si avanza en el conocimiento científico o controla variables de confusión.<sup>15</sup> Carece de pasos explícitos para la formulación de hipótesis o el control experimental.<sup>20</sup>

#### 2.1.2 KDD y SEMMA

El proceso KDD se centra fuertemente en la tubería de extracción y transformación de datos, pasando de los datos en bruto al "conocimiento" mediante la minería de datos.<sup>21</sup> SEMMA (Sample, Explore, Modify, Model, Assess) es una metodología propietaria de SAS que refleja la KDD pero omite la fase crucial de

"Entendimiento de Negocio/Dominio" al principio.<sup>23</sup>

- **Limitaciones científicas:** Ambos marcos tratan la "minería" de patrones como el objetivo terminal. No exigen la falsificación de estos patrones ni la evaluación de su validez causal. En la investigación científica, descubrir un patrón es simplemente la generación de una hipótesis; KDD lo trata como el descubrimiento del conocimiento en sí mismo.<sup>21</sup>

## 2.2 El proceso de Ciencia de Datos en Equipo (TDSP) y adaptaciones modernas

El TDSP de Microsoft intenta modernizar CRISP-DM integrando prácticas ágiles de desarrollo de software y enfatizando la reproducibilidad y la colaboración.<sup>18</sup> Introduce definiciones de roles (por ejemplo, Científico de Datos, Arquitecto de Soluciones) y artefactos estandarizados. Aunque TDSP mejora la *gestión* de la ciencia de datos, sigue tratando el modelo como un producto de software en lugar de un instrumento científico.

## 2.3 Hacia un modelo de procesos científicos: GCRISP-DS y más allá

Para adaptar estos marcos a la investigación científica, los investigadores han propuesto extensiones como el Proceso Estándar Generalizado Interindustrial para la Ciencia de Datos (GCRISP-DS).<sup>26</sup> Este marco aborda explícitamente los "problemas de robustez" y permite interacciones dinámicas entre fases para manejar la complejidad de los datos científicos.

Una metodología verdaderamente científica para la IA debe superponer estos modelos de proceso con la lógica del diseño experimental. La fase de "Modelado" de CRISP-DM debe reconceptualizarse como "Prueba de Hipótesis", y la fase de "Evaluación" debe ampliarse para incluir pruebas de significación estadística, estudios de ablación y análisis causal.

## 2.4 Análisis comparativo de marcos metodológicos

La siguiente tabla contrasta el enfoque industrial con la metodología científica propuesta para la IA, destacando el cambio de la utilidad a la validez (ver Tabla 8).

**Tabla 8: Características de métodos de investigación tradicionales y con IA**

Característica	CRISP-DM / KDD	Método científico	Metodología de IA Científica
Objetivo principal	Despliegue / Valor de negocio	Generación de conocimiento	Descubrimiento robusto de conocimiento mediante computación
Punto	Problema empresarial	Observación / Teoría	Anomalía de dominio / Hipótesis basada en datos
Validación	Umbral métrico (por ejemplo, precisión)	Falsificación experimental	Significación estadística y verificación causal
Control	Limpieza de datos	Grupo de Control Experimental	Estudios de Ablación y Líneas Base Sintéticas
Resultado	Modelo desplegado	Teoría revisada por pares	Modelo reproducible y mecanismo causal

### 3. Generación de hipótesis y sesgo inductivo

En el método científico tradicional, una hipótesis es una explicación tentativa derivada de la teoría u observación, que luego se somete a pruebas empíricas.<sup>27</sup> En IA, el concepto de hipótesis es dual: existe tanto en la formulación humana de la pregunta de investigación como en la formulación matemática del algoritmo de aprendizaje.

#### 3.1 Sesgo inductivo como hipótesis

En aprendizaje automático, el "espacio de hipótesis" se refiere al conjunto de todas las funciones que un algoritmo de aprendizaje puede seleccionar.<sup>14</sup> La selección de un algoritmo específico (por ejemplo, una Red Neuronal Convolutiva sobre un Bosque Aleatorio) impone un *sesgo inductivo*: un conjunto de supuestos sobre la estructura subyacente de los datos.<sup>13</sup>

- **Implicación científica:** Cuando un investigador selecciona una arquitectura de modelo, está formulando implícitamente una hipótesis científica sobre los datos. Por ejemplo, elegir una Red Neuronal Convolutiva (CNN) plantea la hipótesis de que los datos presentan invariancia traslacional y correlaciones espaciales locales.<sup>29</sup>
- **Requisito metodológico:** La investigación rigurosa en IA requiere la articulación explícita de estos sesgos inductivos. La naturaleza de "caja negra" del aprendizaje profundo a menudo oscurece estas suposiciones, dificultando distinguir si el rendimiento se debe a una suposición teórica correcta (el sesgo inductivo) o simplemente a un sobreajuste.<sup>13</sup>

### 3.2 IA generativa y generación automatizada de hipótesis

Los avances recientes han introducido el uso de Grandes Modelos de Lenguaje (LLMs) e IA Generativa para automatizar el propio proceso de generación de hipótesis.<sup>30</sup> sistemas como "The AI Scientist" o "FieldSHIFT" utilizan LLMs para explorar vastos repositorios de literatura científica, identificando conexiones latentes (por ejemplo, entre biología del desarrollo y neurociencia) para proponer hipótesis nuevas y comprobables.<sup>8</sup>

- **Descubrimiento basado en literatura (LBD):** Los agentes de IA pueden realizar "FieldSHIFTing", mapeando conceptos de un dominio a otro para generar hipótesis "fuera de distribución" que los investigadores humanos podrían pasar por alto debido a la compartimentación disciplinaria.<sup>32</sup>
- **Prompting adversarial:** Para garantizar la robustez, los investigadores utilizan prompting adversarial para obligar a los LLM a criticar sus propias hipótesis, exponiendo sesgos y lagunas lógicas antes de las pruebas experimentales.<sup>31</sup>

### 3.3 El ciclo de hipótesis basadas en datos

La metodología de la ciencia impulsada por IA funciona como un ciclo:

1. **Análisis exploratorio de datos (EDA):** El aprendizaje no supervisado (agrupamiento, reducción de dimensionalidad) revela anomalías o estructuras en datos de alta dimensión.<sup>33</sup>
2. **Formulación de hipótesis:** Estas estructuras se formalizan en una hipótesis (por ejemplo, "El Clúster A representa un fenotipo distinto causado por el gen X").
3. **Selección del modelo como especificación de hipótesis:** Un modelo supervisado está diseñado para discriminar esta estructura, codificando la hipótesis en su arquitectura y características.<sup>30</sup>
4. **Verificación:** El rendimiento del modelo se evalúa frente a una hipótesis nula mediante controles estadísticos.

## 4. Diseño experimental en el laboratorio computacional

La validez de cualquier afirmación científica depende del rigor del diseño experimental. En la investigación en IA, el "experimento" es el entrenamiento y evaluación de un modelo. Sin embargo, la naturaleza estocástica de la optimización y la complejidad de las canalizaciones de datos introducen variables de confusión únicas que deben ser controladas.

### 4.1 La necesidad de líneas base

Un fracaso generalizado en la investigación en IA es la falta de referencias rigurosas. No se puede decir que un modelo complejo de aprendizaje profundo sea superior a menos que supere al modelo razonable más simple y al estado de la tecnología (SOTA) bajo condiciones idénticas.<sup>34</sup>

- **Líneas base simples:** La investigación debe comenzar con modelos "ingenuos" (por ejemplo, regresión logística, bosque aleatorio o incluso un clasificador de clase mayoritaria) para establecer la dificultad de la tarea.<sup>35</sup> Si un modelo de Transformer solo supera marginalmente a una regresión lineal, la hipótesis de que "las dependencias secuenciales complejas son críticas" se debilita.
- **Líneas de base sólidas:** Deben compararse con versiones ajustadas de métodos competidores. Utilizar hiperparámetros por defecto como líneas base mientras se

ajusta mucho el modelo propuesto constituye un fallo metodológico.<sup>36</sup>

## 4.2 Variables de control y estudios de ablación

En las ciencias físicas, las variables de control se mantienen constantes para aislar el efecto de la variable independiente.<sup>38</sup> En la IA, los "estudios de ablación" sirven como el mecanismo principal de control.

- **Lógica de la ablación:** Un estudio de ablación elimina sistemáticamente componentes del sistema (por ejemplo, un mecanismo de atención específico, un paso de aumento de datos o un conjunto de características) para medir su contribución marginal al rendimiento.<sup>40</sup>
- **Estándar metodológico:** Un trabajo riguroso sobre IA debe incluir una tabla de ablación. Si una nueva arquitectura propone tres módulos novedosos (A, B y C), el experimento debe probar el sistema solo con A+B, A+C, B+C y el modelo base.<sup>41</sup> Esto evita la falacia del "fregadero de cocina", donde se acepta un modelo complejo a pesar de que solo un componente realmente proporciona beneficio.
- **Inspiración neurocientífica:** Las técnicas modernas de ablación se inspiran en la neurociencia (estudios de lesiones), desactivando selectivamente neuronas o capas para mapear la localización funcional dentro de la red.<sup>43</sup>

## 4.3 Fuga de datos: El factor de confusión de la validez

La fuga de datos es el equivalente computacional a contaminar una muestra. Ocurre cuando la información del conjunto de pruebas (el "futuro") influye inadvertidamente en el proceso de entrenamiento.<sup>45</sup>

- **Tipos de fugas:**
  - **Fuga de preprocesamiento:** Normalizar datos (por ejemplo, calcular media y varianza) usando todo el conjunto de datos *antes* de dividirlos en entrenamiento/prueba. Esto filtra la distribución del conjunto de pruebas hacia el conjunto de entrenamiento.<sup>46</sup>
  - **Fuga temporal:** En la predicción de series temporales, usar datos futuros para predecir el pasado, o barajar aleatoriamente que destruye el orden temporal.<sup>48</sup>
- **Metodología de prevención:** La "Regla de Oro" del diseño experimental de



aprendizaje automático es separar primero los datos y luego aplicar todas las transformaciones (imputación, escalado, ingeniería de características) de forma independiente a los conjuntos de entrenamiento y prueba. Se deben usar pipelines (por ejemplo, scikit-learn Pipelines) para hacer cumplir esta encapsulación.<sup>47</sup>

## 4.4 Datos sintéticos como control experimental

Los datos sintéticos ofrecen una herramienta poderosa para la validación metodológica. Generando datos donde se conocen (y controlan las relaciones causales de la verdad fundamental), los investigadores pueden verificar si su modelo identifica correctamente estas relaciones.<sup>30</sup>

- **Validación del método:** Si un modelo afirma detectar los factores causales de una enfermedad, primero debe demostrarse que detecta los factores en un conjunto de datos sintético donde el mecanismo de la enfermedad está definido matemáticamente.<sup>50</sup>
- **Privacidad y control de sesgos:** Los datos sintéticos también pueden utilizarse para crear conjuntos de datos equilibrados que controlan el sesgo demográfico, permitiendo aislar variables específicas (por ejemplo, raza, género) para comprobar la equidad algorítmica.<sup>51</sup>

## 5. Rigor estadístico y evaluación del modelo

El simple hecho de reportar cifras de mayor precisión es insuficiente para la prueba científica. La variabilidad en el entrenamiento de aprendizaje automático (debido a semillas aleatorias, inicialización de pesos y baraja de datos) significa que las diferencias de rendimiento a menudo pueden atribuirse al azar.<sup>53</sup>

### 5.1 Pruebas de hipótesis para la comparación de modelos

Para afirmar que el Modelo A es superior al Modelo B, los investigadores deben rechazar la hipótesis nula de que sus distribuciones de rendimiento son idénticas.

- **La insuficiencia de las pruebas t ingenuas:** Las pruebas t estándar de estudiantes suelen asumir muestras independientes, lo cual se viola en la validación cruzada

cuando los conjuntos de entrenamiento se solapan.<sup>53</sup>

- **Prueba de McNemar:** Para comparar dos clasificadores en un solo conjunto de pruebas, la prueba de McNemar es el estándar. Evalúa la tabla de contingencia de predicciones correctas/incorrectas, centrándose en los casos en los que los modelos *discrepan*.<sup>56</sup>
- **Prueba t de validación cruzada 5x2:** Para una comparación más robusta que involucre la variabilidad del entrenamiento, Dietterich recomienda la prueba t pareada de validación cruzada 5x2. Esto implica repetir la validación cruzada de 2 veces cinco veces, proporcionando una mejor estimación de la varianza que la CV estándar de 10 veces.<sup>53</sup>

## 5.2 Intervalos de confianza (IC)

Informar de una estimación puntual (por ejemplo, "Precisión: 94,5%") es científicamente engañoso. Las metodologías deben requerir la notificación de Intervalos de Confianza (IC) para expresar la incertidumbre de la estimación.<sup>59</sup>

- **Métodos de arranque:** El bootstrapping (remuestreo del conjunto de pruebas con reemplazo) permite el cálculo empírico de ICs para cualquier métrica (precisión, F1, AUC) sin asumir una distribución normal.<sup>59, 60, 61</sup>
- **Estándar de reporte:** Los resultados deben informarse como  $\mu \pm \sigma$  o con un IC del 95% (por ejemplo, [93,2%, 95,8%]). Si los IC de la línea base y del modelo propuesto se solapan, la mejora no es estadísticamente significativa.<sup>59, 62</sup>

## 5.3 Pruebas de Esfuerzo Adversariales

Así como las estructuras de ingeniería se someten a pruebas de resistencia, los modelos de aprendizaje automático deben someterse a una evaluación adversarial. Una alta precisión media puede ocultar una fragilidad catastrófica en los peores escenarios.<sup>31</sup>

- **Ejemplos adversariales:** Entradas perturbadas intencionadamente (a menudo imperceptiblemente para los humanos) para inducir error del modelo. Probar con estos ejemplares evalúa la robustez y estabilidad inductiva del modelo (ver Tabla 9).<sup>63</sup>
- **Utilidad científica:** Las pruebas adversariales sirven como mecanismo de

falsificación de la hipótesis de que el modelo ha aprendido el concepto "verdadero". Si un "panda" se clasifica como "gibón" tras una perturbación menor del ruido de píxel, se refuta la hipótesis de que el modelo reconoce a los pandas por sus rasgos visuales.<sup>64</sup>

**Tabla 9: Métodos de evaluación estadística**

Método de prueba	Caso de uso	Suposición estadística	Limitación
<b>Test de McNemar</b>	Comparando 2 clasificadores en un solo conjunto de pruebas.	Datos nominales emparejados.	Ignora la variación del conjunto de entrenamiento.
<b>Prueba t de CV 5x2</b>	Comparando algoritmos entre divisiones de conjuntos de datos.	Distancia aproximadamente normal de diferencias.	Computacionalmente caro (10 entrenamientos).
<b>Bootstrap CI</b>	Estimar la incertidumbre de una métrica.	El distrito remuestreado aproxima la población.	Puede estar sesgado en muestras pequeñas (la regla .632+ corrige esto).
<b>Prueba adversarial</b>	Análisis de robustez / peor caso.	Ninguno (basado en optimización).	Específico del método de ataque (por ejemplo, FGSM, PGD).

## 6. Reproducibilidad, transparencia y pre-registro

La "Crisis de Reproducibilidad" en la IA refleja la de la psicología y la medicina, agravada por datos propietarios, algoritmos estocásticos y el "hacking de ajuste".<sup>45</sup> Una metodología científica para la IA exige una estricta adhesión a los estándares de reproducibilidad.

## 6.1 La Lista de Verificación de Reproducibilidad

Propuesta por Pineau et al. y adoptada por las principales conferencias (NeurIPS, ICML), la Lista de Verificación de Reproducibilidad convierte las mejores prácticas en requisitos obligatorios de notificación.<sup>66</sup>

- **Elementos clave:**
  - **Datos:** Enlaces a conjuntos de datos descargables, descripciones detalladas del preprocesamiento y divisiones de entrenamiento/prueba.<sup>67</sup>
  - **Código:** Especificación de dependencias (contenedores Docker, entornos Conda), semillas aleatorias y comandos exactos para reproducir resultados.<sup>67</sup>
  - **Hiperparámetros:** Divulgación completa del rango de hiperparámetros buscados y del método de selección (por ejemplo, búsqueda en cuadrilla, optimización bayesiana).<sup>68</sup> Esto previene el fenómeno de la "semilla de la suerte", donde solo se reporta la mejor carrera.

## 6.2 Pre-registro de la investigación

Para combatir el sesgo de publicación y el "p-hacking" (ajustar el modelo hasta encontrar resultados significativos), la comunidad de ML se está moviendo hacia la Pre-Inscripción.<sup>69</sup>

- **Mecanismo:** Los investigadores presentan un artículo que contiene la Introducción, el Trabajo Relacionado y *el Plan Experimental* (incluyendo conjuntos de datos, métricas e hipótesis) *antes* de ejecutar los experimentos.
- **Beneficio científico:** Esto desvincula la evaluación de la *idea* del *resultado*. Garantiza que se publiquen resultados negativos (por ejemplo, "Esta arquitectura no mejoró el rendimiento"), evitando el problema del cajón de archivos y evitando que la comunidad persiga callejones sin salida.<sup>70, 71</sup>
- **Plataformas:** El Open Science Framework (OSF) proporciona plantillas para pre-registrar estudios observacionales y experimentales, que pueden adaptarse para la investigación en aprendizaje automático.<sup>72</sup>

## 6.3 Control de versiones para datos (DVC)

El rigor científico requiere que los *datos* se versionen con la misma granularidad que el código. Herramientas como el Control de Versiones de Datos

(DVC) permiten a los investigadores consultar la versión exacta del conjunto de datos utilizado para un experimento específico, asegurando que los cambios en la cadena de datos no confundan los resultados.<sup>65, 73</sup>

## 7. Causalidad y explicación: más allá de la correlación

La limitación epistemológica más significativa del aprendizaje automático estándar es su dependencia de la correlación. Para lograr la comprensión científica, la metodología debe ascender la "Escalera de Causalidad" desde la Asociación ( $P(y|x)$ ) hasta la Intervención ( $P(y|do(x))$ ) y los Contrafactuales.<sup>74</sup>

### 7.1 Inferencia causal vs. modelado predictivo

- **Modelado predictivo:** Pregunta "¿Cuál es el valor probable de Y dado X?" Es tolerante a correlaciones espurias (por ejemplo, prediciendo lluvia procedente de paraguas).<sup>75, 76</sup>
- **Inferencia causal:** Pregunta "¿Qué pasará con Y si cambio X?" Esto requiere un modelo causal estructural (SCM) o un grafo acíclico dirigido (DAG) para codificar supuestos sobre la confusión y la direccionalidad.<sup>77, 78, 79</sup>
- **Integración:** La metodología de "ML causal" utiliza aprendizaje automático para estimar efectos causales (por ejemplo, Doble Aprendizaje Automático) o para descubrir grafos causales a partir de datos. Esto es fundamental en campos como la medicina personalizada, donde el objetivo es la intervención (tratamiento), no solo el pronóstico.<sup>80</sup>

### 7.2 IA explicable (XAI) como falsificación científica

La IA explicable a menudo se presenta como un mecanismo de confianza, pero científicamente, es una herramienta de falsificación.<sup>81, 82</sup>

- **Puntos de vista divergentes vs. convergentes:**
  - **Vista convergente:** La explicación confirma que el modelo se basa en características conocidas del dominio (por ejemplo, un clasificador tumoral que observa el tumor). Esto genera confianza.<sup>83</sup>

- **Visión divergente:** La explicación revela que el modelo utiliza características inesperadas (por ejemplo, la etiqueta del escáner en la radiografía). Esta *divergencia* refuta la hipótesis de que el modelo ha aprendido la patología, lo que genera nuevas preguntas científicas o corrección del modelo.<sup>84</sup>
- **Riesgos de racionalización post-hoc:** Métodos como LIME y SHAP proporcionan aproximaciones del comportamiento del modelo. Los investigadores deben validar que estas explicaciones son fieles al modelo y no "placebicas" (convincientes pero erróneas).<sup>85</sup> El uso de la interpretabilidad mecanicista—ingeniería inversa de los pesos y activaciones (por ejemplo, IA "microscopio")—ofrece un camino más riguroso para la explicación que la importancia superficial de las características.<sup>86, 87</sup>

La aplicación del método científico a la Inteligencia Artificial transforma la disciplina de una alquimia de parámetros en una rigurosa ciencia de la inteligencia (ver Tabla 10). Esta transición requiere un marco metodológico holístico:<sup>88, 89</sup>

1. **Proceso:** Adoptar modelos científicos de procesos (como TDSP/GCRISP-DS modificados) que prioricen la prueba de hipótesis sobre métricas puras.
2. **Diseño:** Implementación de controles rigurosos mediante estudios de ablación, datos sintéticos y estricta prevención de fugas de datos.
3. **Validación:** Ir más allá de la precisión hacia pruebas de significación estadística (McNemar's, Bootstrap CIs) y pruebas de esfuerzo adversariales.
4. **Reproducibilidad:** Institucionalizar listas de pre-registro y de comprobación de artefactos para asegurar que los hallazgos sean acumulativos y verificables.
5. **Explicación:** Elevar la inferencia causal y XAI para distinguir mecanismos causales robustos de correlaciones frágiles.

**Tabla 10: Recomendaciones metodológicas clave para investigadores en IA**

Dominio	Recomendación	Herramienta/Acción metodológica
Hipótesis	Explícitamente declara sesgos inductivos.	<b>Pre-registro</b> de arquitectura e hipótesis en OSF.

<b>Diseño</b>	Establece la dificultad y la contribución.	<b>Líneas de base simples</b> (por ejemplo, regulación logística) y <b>estudios de ablación</b> .
<b>Datos</b>	Evita fugas y asegúrate de ser independiente.	<b>Encapsulación de tubería</b> y <b>división de series temporales</b> .
<b>Evaluación</b>	Cuantifica la incertidumbre y la importancia.	<b>Intervalos de confianza de McNemar's Test &amp; Bootstrap</b> .
<b>Reproducibilidad</b>	Asegurar la replicabilidad de los resultados.	<b>Lista de verificación</b> de Pineau, <b>Docker</b> y <b>fijación con semillas aleatorias</b> .
<b>Causalidad</b>	Distingue la predicción de la causa.	<b>Grafos causales (DAGs)</b> y biblioteca <b>DoWhy</b> para refutación.
<b>Robustez</b>	Prueba el peor de los casos.	<b>Generación de ataques adversariales</b> y <b>pruebas de estrés con datos sintéticos</b> .

# Conclusión

No se puede ignorar que la ética en IA ha dejado de ser una consideración filosófica externa para convertirse en una variable técnica interna. Los sesgos algorítmicos, la discriminación y la falta de equidad no son solo "efectos secundarios" sociales, sino indicadores de una metodología de investigación defectuosa: muestreo sesgado, etiquetas ruidosas, funciones de pérdida mal definidas o validación en subgrupos insuficientes.

A pesar de la existencia de marcos regulatorios emergentes como la *EU AI Act* o las recomendaciones de la UNESCO, la investigación académica a menudo carece de mecanismos formales para integrar estas consideraciones en el diseño experimental. La documentación de los datasets es frecuentemente inexistente o superficial; los investigadores raramente reportan la procedencia de los datos, el consentimiento de los sujetos o las limitaciones de uso, lo que perpetúa la creación de modelos "tóxicos" o legales pero ilegítimos.<sup>32</sup> La falta de adopción de herramientas estandarizadas como *Datasheets for Datasets* o *Model Cards* en la etapa de investigación (y no solo en despliegue) es una brecha crítica que este libro busca cerrar.<sup>34</sup>

La evidencia vislumbra un problema de capital humano. Los programas académicos y bootcamps de ciencia de datos se centran desproporcionadamente en la sintaxis de la programación y la arquitectura de modelos, descuidando la formación en diseño de investigación, pensamiento crítico y metodología científica. Esto resulta en una generación de investigadores noveles que pueden implementar una red neuronal compleja pero que carecen de las herramientas para formular una pregunta de investigación novedosa, diseñar una estrategia de validación que controle variables confusoras o escribir un *paper* que sobreviva a una revisión por pares rigurosa.

Además, la divergencia de incentivos entre la academia ("publicar o perecer")



y la industria ("retorno económico o morir") crea una esquizofrenia metodológica. Mientras la academia premia la novedad teórica a menudo sobre problemas de juguete, la industria necesita soluciones robustas y mantenibles sobre datos sucios y cambiantes. Por ende, se abordó cómo aplicar el método científico en entornos corporativos donde los tiempos son cortos y los datos son propietarios, adaptando el rigor académico a la realidad pragmática del negocio.

En respuesta a la problemática multidimensional descrita, el libro *"Metodología de la investigación científica aplicada a la inteligencia artificial y la ciencia de datos"* se define como un compendio normativo y práctico. Su alcance trasciende la enseñanza de herramientas técnicas para centrarse en la estructuración del pensamiento científico y el proceso de investigación. La obra se posiciona en la intersección de la filosofía de la ciencia, la ingeniería de software empírica y la estadística aplicada.

En conclusión, se argumentó que la IA y la ciencia de datos no son meramente ciencias naturales (descubrimiento de patrones preexistentes) ni puramente matemáticas (demostración de teoremas), sino ciencias de diseño que estudian artefactos creados por el ser humano para cumplir objetivos en entornos inciertos. Por ende, recomendamos emplear esta obra como guía sobre cómo documentar la motivación, composición, proceso de recolección, preprocesamiento y usos recomendados de los datos. Esto responde directamente a los problemas de transparencia ética y técnica.

# Bibliografía

1. Hey, T. (2012). The Fourth Paradigm – Data-Intensive Scientific Discovery. In: Kurbanoglu, S., Al, U., Erdogan, P.L., Tonta, Y., Ucak, N. (eds) E-Science and Information Management. IMCW 2012. Communications in Computer and Information Science, vol 317. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-33299-9\\_1](https://doi.org/10.1007/978-3-642-33299-9_1)
2. Choque Gutiérrez, J. J., Mamani Figueroa, J. D., Puita Choque, G. A., Castro Angulo, D. D., & Choque Matos, J. (2025). Ciencia Abierta y Colaborativa en la Era de la Inteligencia Artificial. *Revista Veritas De Difusão Científica*, 6(1), 2162–2172. <https://doi.org/10.61616/rvdc.v6i1.505>
3. Shen, L., Bai, J., Wang, J., & Shen, B. (2021). The fourth scientific discovery paradigm for precision medicine and healthcare: Challenges ahead. *Precision clinical medicine*, 4(2), 80–84. <https://doi.org/10.1093/pcmedi/pbab007>
4. Carlin, J. B. and Moreno-Betancur, M. (2025). Rejoinder to commentaries on: on the uses and abuses of regression models: a call for reform of statistical practice and teaching. *Statistics in Medicine*, 44(13-14). <https://doi.org/10.1002/sim.70065>
5. Galit, S. (2025). To Explain or to Predict?. *Statist. Sci.*, 25 (3) 289-310. <https://doi.org/10.1214/10-STS330>
6. Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *The British journal of mathematical and statistical psychology*, 66(1), 8–38. <https://doi.org/10.1111/j.2044-8317.2011.02037.x>

7. Guersenzvaig, A. (2024). Can machine learning make naturalism about health truly naturalistic? A reflection on a data-driven concept of health. *Ethics Inf Technol*, 26(2). <https://doi.org/10.1007/s10676-023-09734-6>
8. Chubb, J., Cowling, P. y Reed, D. (2022). Acelerando para mantenerse al día: explorando el uso de la IA en el proceso de investigación. *AI & Soc.*, 37, 1439-1457. <https://doi.org/10.1007/s00146-021-01259-0>
9. Qadri, YA, Shaikh, S., Ahmad, K., Choi, I., Kim, SW y Vasilakos, AV (2025). Inteligencia artificial explicable: Una perspectiva sobre el descubrimiento de fármacos. *Pharmaceutics* , 17 (9), 1119. <https://doi.org/10.3390/pharmaceutics17091119>
10. Mazzocchi F. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO reports*, 16(10), 1250–1255. <https://doi.org/10.15252/embr.201541001>
11. Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810–821. <https://doi.org/10.1086/684083>
12. Emden, M. (2021). ¿Reintroduciendo el método científico para introducir *la investigación científica* en las escuelas? *Sci & Educ*, 30, 1037–1073. <https://doi.org/10.1007/s11191-021-00235-w>
13. Musslick, S., Bartlett, L.K., Chandramouli, S.H. Dubova, M., Gobet, F., Griffiths, T.L., Hullman, J., Rey, R.D., Kutz, J., Lucas, D.G., Mahesh, S., Pestilli, F., Sloman, S.J., & Holmes, O.R. (2025). Automatización de la práctica científica: oportunidades, desafíos e implicaciones. *Proc. Natl. Sci.* 122(5), e2401238121. <https://doi.org/10.1073/pnas.2401238121>

14. Chen, M., Huang, Y., Wang, W. *et al.* (2024). Model inductive bias enhanced deep reinforcement learning for robot navigation in crowded environments. *Complex Intell. Syst.* 10, 6965–6982. <https://doi.org/10.1007/s40747-024-01493-1>
15. Massahiro, A., Cordeiro, R., and Goldman, A. (2025). Integrating CRISP-DM and eXtreme Programming in Agile Data Science Projects. *European Journal of Electrical Engineering and Computer Science*, 9(4), 1-7. <https://doi.org/10.24018/ejece.2025.9.4.739>
16. Nguyen, G., Dlugolinsky, S., Bobák, M. *et al.* (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artif Intell Rev*, 52, 77–124. <https://doi.org/10.1007/s10462-018-09679-z>
17. Benganga, J., Kukuni, T., Kotze, B. y Lenkoe, L. (2025). Análisis comparativo de la eficiencia de modelos basado en algoritmos disímiles para el aprendizaje y la corrección de imágenes como método de detección de fallos. *Matemáticas*, 13 (11), 1835. <https://doi.org/10.3390/math13111835>
18. Paik, I. (2024). Análisis automático de big data mediante composición de servicios basada en IA para ciudades inteligentes. En: Murakami, Y., Kimura, K. (eds.) *Computación de servicios centrada en el ser humano para ciudades inteligentes*. Springer, Singapur. [https://doi.org/10.1007/978-981-97-0779-9\\_4](https://doi.org/10.1007/978-981-97-0779-9_4)
19. Durango Vanegas, C.E., Giraldo Mejía, J.C., Vargas Agudelo, F.A., & Soto Durán, D.E. (2023). A Representation Based on Essence for the CRISP-DM Methodology. *Computación y Sistemas*, 27(3), 675-689.

<https://doi.org/10.13053/cys-27-3-3446>

20. Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: a systematic literature review. *PeerJ. Computer science*, 6, e267. <https://doi.org/10.7717/peerj-cs.267>
21. Cios, K.J., Pedrycz, W., Swiniarski, R.W. (1998). Data Mining and Knowledge Discovery. In: Data Mining Methods for Knowledge Discovery. The Springer International Series in Engineering and Computer Science, vol 458. Springer, Boston, MA. [https://doi.org/10.1007/978-1-4615-5589-6\\_1](https://doi.org/10.1007/978-1-4615-5589-6_1)
22. Pujar, S., Zheng, Y., Buratti, L. *et al.* (2024). Analyzing source code vulnerabilities in the D2A dataset with ML ensembles and C-BERT. *Empir Software Eng*, 29(48). <https://doi.org/10.1007/s10664-023-10405-9>
23. Hochkamp, F., Scheidler, AA y Rabe, M. (2025). Revisión de modelos de madurez para minería de datos y propuesta de un prototipo de modelo de madurez de preparación de datos para minería de datos. *Computers* , 14 (4), 146. <https://doi.org/10.3390/computers14040146>
24. de Mast, J., Lokkerbol, J. (2024). Diagramas DAPS para la definición de proyectos de ciencia de datos. *J Big Data*, 11(50). <https://doi.org/10.1186/s40537-024-00916-7>
25. Shaheen, F.A., Gul, A., Ganai, N. *et al.* (2025). Deep learning-enabled cherry price forecasting and real-time system deployment across multi-market supply chains in India. *Sci Rep*. <https://doi.org/10.1038/s41598-025-30980-9>
26. Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & Emmert-Streib, F. (2021). Ensuring the Robustness and Reliability of

- Data-Driven Knowledge Discovery Models in Production and Manufacturing. *Frontiers in artificial intelligence*, 4, 576892. <https://doi.org/10.3389/frai.2021.576892>
27. Daphne, E., and Whitaker, K. (2019). Point of View: Data science for the scientific life cycle. *eLife*, 8, e43979. <https://doi.org/10.7554/eLife.43979>
  28. Li, J. J., & Tong, X. (2020). Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines. *Patterns (New York, N.Y.)*, 1(7), 100115. <https://doi.org/10.1016/j.patter.2020.100115>
  29. Butz, M.V. (2021). Hacia una IA fuerte. *Künstl Intell*, 35, 91-101. <https://doi.org/10.1007/s13218-021-00705-x>
  30. Xianyu, Z., Correia, C., Ung, C. Y., Zhu, S., Billadeau, D. D., & Li, H. (2024). The Rise of Hypothesis-Driven Artificial Intelligence in Oncology. *Cancers*, 16(4), 822. <https://doi.org/10.3390/cancers16040822>
  31. Zhang, Y., Khan, S.A., Mahmud, A. *et al.* Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artif. Intell.* 1, 14 (2025). <https://doi.org/10.1038/s44387-025-00019-5>
  32. Mapstone, C., & Plusa, B. (2025). Machine learning approaches for image classification in developmental biology and clinical embryology. *Development (Cambridge, England)*, 152(4), DEV202066. <https://doi.org/10.1242/dev.202066>
  33. Najm, IA, Hamoud, AK, Lloret, J. y Bosch, I. (2019). Enfoque de predicción mediante aprendizaje automático para mejorar el control de la congestión en entornos IoT 5G. *Electronics* , 8 (6), 607.

<https://doi.org/10.3390/electronics8060607>

34. Almeida, A. P., Santana Júnior, C. A. de, & Brito, T. H. da S. (2023). LOS DILEMAS ÉTICOS DE LA INTELIGENCIA ARTIFICIAL EN LOS SERVICIOS DE INFORMACIÓN BIBLIOTECARIA. *Revista EDICIC*, 3(3).  
<https://doi.org/10.62758/re.v3i3.263>
35. Almeida-Campos, S. (2024). Open science and artificial intelligence in Revista Médica Electrónica | La ciencia abierta y la inteligencia artificial en la Revista Médica Electrónica. *Revista Medica Electronica*, 46(1)
36. Choque Gutiérrez, J. J., Mamani Figueroa, J. D., Puita Choque, G. A., Castro Angulo, D. D., & Choque Matos, J. (2025). Ciencia Abierta y Colaborativa en la Era de la Inteligencia Artificial. *Revista Veritas de Difusão Científica*, 6(1). <https://doi.org/10.61616/rvdc.v6i1.505>
37. de Quijada, P. E. S., Rodríguez, M. A. M., Lingan, A. M. A., Ñaca, P. Q., & Zamora, M. C. J. (2025). DESAFÍO DE LA INTELIGENCIA ARTIFICIAL EN EL PROCESO DE INVESTIGACIÓN Y PUBLICACIÓN CIENTÍFICA: UNA REVISIÓN SISTEMÁTICA DE LA LITERATURA DE AMÉRICA LATINA. *LUMEN ET VIRTUS*, 16(45).  
<https://doi.org/10.56238/levv16n45-043>
38. Gil Gamboa, K. de los A., & Gaibor Vera, F. M. (2025). Ética en la investigación científica: desafíos actuales y perspectivas futuras. *Conexión Científica Revista Internacional*, 2(4). <https://doi.org/10.71068/7df64t58>
39. Guevara-Pezoa, F. (2023). La ciencia abierta y su relación con la innovación: una revisión bibliométrica. *Investigación Bibliotecológica*:

40. Lira Beltrán, R. A., Orizaga Trejo, J. A., Castañeda González, C. A., & Cruz Herrera, H. (2024). La aplicación de la Inteligencia Artificial en la inserción productiva de egresados universitarios. *TIES, Revista de Tecnología e Innovación En Educación Superior*, 10. <https://doi.org/10.22201/dgtic.26832968e.2024.10.14>
41. López-Bermúdez, F. L., Hidalgo-Hidalgo, W. A., Medrano-Freire, E. L., & Barba-López, R. A. (2024). Las matemáticas aplicadas como herramienta para la resolución de problemas de la ciencia y la sociedad. *MQRInvestigar*, 8(4). <https://doi.org/10.56048/mqr20225.8.4.2024.7408-7421>
42. Lopezosa, C. (2023). ChatGPT y comunicación científica: hacia un uso de la Inteligencia Artificial que sea tan útil como responsable. *Hipertext.Net*, 26. <https://doi.org/10.31009/hipertext.net.2023.i26.03>
43. Maisincho Camacho, J. D. (2025). Análisis del uso de inteligencia artificial y ciencia de datos en la optimización de procesos de investigación científica en el nivel académico. *Revista Multidisciplinar Ciencia y Descubrimiento*, 3(2). <https://doi.org/10.63816/4qsebs84rcd>
44. Mullo-Romero, E. E. del C., Brusela Vásquez-Farfán, N. I., & Ramiro Chávez-Escobar III, H. (2024). Inteligencia Artificial Aplicada al Sector Turístico: Evolución y Tendencias de Investigación. *Polo Del Conocimiento*, 9(11)



45. Pavajeau Hernández, J. J., Díaz Alay, Z. S., Alvarado Chicaiza, E. H., & Mora Solórzano, S. J. (2024). Aportes de la filosofía contemporánea en el desarrollo de la investigación científica. *Revista Social Fronteriza*, 4(3). [https://doi.org/10.59814/resofro.2024.4\(3\)275](https://doi.org/10.59814/resofro.2024.4(3)275)
46. Rea-Zurita, J., Luna-Montalvo, A., & Aldaz-Yépez, I. (2024). Divulgación Científica en los Medios Digitales de Ecuador: Coberturas de Inteligencia Artificial en El Universo y Primicias. *Revista Enfoques de La Comunicación*, 12
47. Rincon, I., Soledispa, B., Sumba, R., Burbano, Z. del C., & Jiménez, F. (2023). Abordajes metodológicos y paradigmas en la investigación científica y tecnológica: una revisión bibliométrica. *Bibliotecas. Anales de Investigación*, 19(1)
48. Tapullima-Mori, C., Mamani-Benito, O., Turpo-Chaparro, J. E., Olivas-Ugarte, L. O., & Carranza-Esteban, R. F. (2024). Inteligencia artificial en la educación universitaria: Revisión bibliométrica en Scopus y Web of Science. *Revista Electrónica Educare*, 28(S). <https://doi.org/10.15359/ree.28-s.18489>
49. Ticona Salluca, H., Borda Colque, J. P., Canqui Flores, B., Yupanqui Bendita, C. E., Hancoo Quispe, J. K., & Torres-Cruz, F. (2023). APLICACIONES DE INTELIGENCIA ARTIFICIAL APLICADA A LA AGRICULTURA PERUANA. *Pensamiento Crítico En La Investigación Científica y Académica*, 21(1)
50. Abdalsalam, M., & Szłapczyńska, J. (2025). Towards Improved Ship Weather Routing Through Multi-Objective Optimization with High

- Performance Computing Support. *TransNav*, 19(1).  
<https://doi.org/10.12716/1001.19.01.12>
51. Abu Sharshouh, A. (2025). The Use of Artificial Intelligence in Accounting and Auditing TT - Muhasebe ve Denetim Alanında Yapay Zekâ Kullanımı. *Karadeniz Ekonomi Araştırmaları Dergisi*, 6(1)
  52. Amruzziyad, Sahwi, & Fitriani, M. I. (2025). Outsourcing Dalam Manajemen Sumber Daya Manusia. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 10(02)
  53. Aviles-Castillo, F., Buele, J., & Palacios-Navarro, G. (2025). Virtual Reality and User Experience: Current Trends and Future Challenges. In *IEEE Access* (Vol. 13). <https://doi.org/10.1109/ACCESS.2025.3554434>
  54. Braza Delgado, R. (2025). Inteligencia artificial y comunicación corporativa: una evolución estructural analizada desde la bibliometría. *ADResearch ESIC International Journal of Communication Research*, 33. <https://doi.org/10.7263/adresic-33-303>
  55. Cricchio, J. (2025). Balancing openness and ownership: open innovation strategies for AI development. *European Journal of Innovation Management*. <https://doi.org/10.1108/EJIM-04-2024-0470>
  56. Cunha, K. C. T., Martens, C. D. P., & Marcolin, C. B. (2025). Mining textual fields from patent documents: systematic review. *Management Review Quarterly*. <https://doi.org/10.1007/s11301-025-00555-z>
  57. de León Pérez, D., Avila-Velazquez, D. I., Macian-Sorribes, H., Salazar-Galán, S., Pulido-Velazquez, M., & Francés García, F. (2025). A

Framework for Enhancing Seasonal Hydrological Forecasting in the Jucar River Basin (Spain). *EGU General Assembly 2025*

58. Fil, N., Slisarenko, R., Deineko, Z., & Morozova, L. (2025). Trends in Artificial Intelligence Research on Education: Topic Modeling Using Latent Dirichlet Allocation. *Bulletin of Kharkov National Automobile and Highway University*, 108. <https://doi.org/10.30977/bul.2219-5548.2025.108.0.17>
59. Frolov, I. E., & Kiselev, V. N. (2025). Artificial Intelligence as a Driver of Breakthrough Technologies: Global Trends and Lessons for Russia. *Studies on Russian Economic Development*, 36(3). <https://doi.org/10.1134/S1075700725700108>
60. Ghisletta, P. (2025). Challenges and opportunities of psychological aging research. In *European Journal of Ageing* (Vol. 22, Issue 1). <https://doi.org/10.1007/s10433-025-00891-9>
61. Gutierrez Mora, I., Hernández Cázares, A. S., Hidalgo Contreras, J. V., López Ayala, J. L., & Velasco Velasco, J. (2025). APLICACIÓN DE TELEDETECCIÓN EN LA ESTIMACIÓN DE RENDIMIENTOS EN CULTIVOS AGRÍCOLAS: UNA REVISIÓN BIBLIOMETRICA. *Tropical and Subtropical Agroecosystems*, 28(1). <https://doi.org/10.56369/tsaes.5613>
62. Gutiérrez-Mora, I., Hernández-Cázares, A. S., Hidalgo-Contreras, J. V., Lopez-Ayala, J. L., & Velasco-Velasco, J. (2025). APPLICATION OF REMOTE SENSING IN THE ESTIMATION OF AGRICULTURAL CROPS YIELDS: A BIBLIOMETRIC REVIEW. In *Tropical and Subtropical Agroecosystems* (Vol. 28, Issue 1). <https://doi.org/10.56369/tsaes.5613>

63. Gutiérrez-Mora, I., Hernández-Cázares, A. S., Hidalgo-Contreras, J. V., López-Ayala, J. L., & Velasco-Velasco, J. (2025). APPLICATION OF REMOTE SENSING IN THE ESTIMATION OF AGRICULTURAL CROPS YIELDS: A BIBLIOMETRIC REVIEW; APLICACIÓN DE TELEDETECCIÓN EN LA ESTIMACIÓN DE RENDIMIENTOS EN CULTIVOS AGRÍCOLAS: UNA REVISIÓN BIBLIOMETRICA. *Tropical and Subtropical Agroecosystems*, 28(1)
64. Harahulia, S., & Symonenko, T. (2025). RESEARCH SUPPORT SERVICES IN DIGITAL HUMANITIES: APPROACHES, TECHNOLOGIES, TOOLS. *Manuscript and Book Heritage of Ukraine*, 2025(2). <https://doi.org/10.15407/rksu.37.189>
65. ILIEV, B. (2025). Furniture design in facilities for preschool education as a basis for healthy children's growth and development. *Repozitorij.Sumfak.Unizg.Hr*.
66. Kashani, M., & Dastani, M. (2025). Evolution of artificial intelligence in medical sciences: a comprehensive scientometrics analysis. *Global Knowledge, Memory and Communication*. <https://doi.org/10.1108/GKMC-09-2024-0586>
67. Linares Rodríguez, E., Linares Rodríguez, E., & Linares Herrera, M. P. (2025). Knowledge management applied to the hydraulic sector in the 21st century. *European Public and Social Innovation Review*, 10. <https://doi.org/10.31637/epsir-2025-1328>
68. Luppichini, M., Capolongo, D., Scardino, G., Scicchitano, G., & Bini, M. (2025). Artificial Intelligence in Geomorphology: A Bibliometric Analysis

- of Trends, Techniques, and Global Research Patterns. *Geosciences (Switzerland)*, 15(9). <https://doi.org/10.3390/geosciences15090331>
69. Nazari, M., Ansari, M., & Pour, M. J. A. (2025). Bibliographic Analysis and Future Research Directions in the Field of Sustainable and Innovative Business Models. *Scientometrics Research Journal*, 11(2). <https://doi.org/10.22070/rsci.2024.19650.1759>
  70. Orlov, A. (2025). Scientific Revolution in Mathematical Research Methods and Artificial Intelligence. *Ergodesign*, 2025(2). <https://doi.org/10.30987/2658-4026-2025-2-160-170>
  71. Parra-Domínguez, J., Sanz Martín, L., López Pérez, G., & Zafra Gómez, J. L. (2025). The disruption of blockchain technology in accounting: a review of scientific progress. In *Journal of Accounting and Organizational Change*. <https://doi.org/10.1108/JAOC-10-2024-0327>
  72. Qiu, B. (2025). Optimization design and application of artificial intelligence in intelligent transportation system. *Proceedings of 2025 6th International Conference on Computer Information and Big Data Applications, CIBDA 2025*. <https://doi.org/10.1145/3746709.3746969>
  73. Silaban, R., Lespiani, I., Daulay, A., Ginting, E., & Sitorus, M. (2025). Penyediaan Bahan Ajar Inovatif Hidrokarbon Berbasis Problem Based-Learning Terintegrasi Kearifan Lokal Minyak Karo. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 10(3)
  74. Taha, A., Khawaja, S., Qureshi, F., & Wahsheh, F. R. (2026). Project portfolio management in the age of artificial intelligence: A review of

- challenges, key features, and future research directions. *Journal of Project Management*, 11
75. Teixeira, M. J., Barbosa, D. J., Dinis-Oliveira, R. J., & Freitas, A. R. (2025). Redefining postmortem interval estimation: the need for evidence-based research to bridge science and justice. In *Frontiers in Microbiology* (Vol. 16). <https://doi.org/10.3389/fmicb.2025.1646907>
  76. ЦИМБАЛ, С., МОВЧАН, О., & ГАПОНОВА, В. (2025). КОРПУСНІ МЕТОДИ У ПЕРЕКЛАДОЗНАВЧИХ ДОСЛІДЖЕННЯХ: СУЧАСНІ ПІДХОДИ ТА НАПРЯМКИ РОЗВИТКУ. *Current Issues of Linguistics and Translation Studies*, 34. <https://doi.org/10.31891/2415-7929-2025-34-14>
  77. Afreen, S., Krohannon, A., Purkayastha, S., & Janga, S. C. (2025). Datawiz-IN: fostering representative innovation in health data science—outcomes from a summer research experience. *BMC Medical Education*, 25(1). <https://doi.org/10.1186/s12909-025-07298-1>
  78. Chen, H., Wang, Y., Li, Y., Lee, Y., Petri, A., & Cha, T. (2023). Computer science and non-computer science faculty members' perception on teaching data science via an experiential learning platform. *Education and Information Technologies*, 28(4). <https://doi.org/10.1007/s10639-022-11326-8>
  79. Darji, A. (2024). Revolutionizing Cancer Research with AI: Health Data Science Lab, UT Arlington. *XRDS: Crossroads, The ACM Magazine for Students*, 30(4). <https://doi.org/10.1145/3665604>
  80. Dwivedi, Y. K., Sharma, A., Rana, N. P., Giannakis, M., Goel, P., & Dutot, V. (2023). Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future

- directions. *Technological Forecasting and Social Change*, 192.  
<https://doi.org/10.1016/j.techfore.2023.122579>
81. Fahd, K., & Miah, S. J. (2023). Designing and evaluating a big data analytics approach for predicting students' success factors. *Journal of Big Data*, 10(1). <https://doi.org/10.1186/s40537-023-00835-z>
  82. Federico, C. A., & Trotsyuk, A. A. (2024). Biomedical Data Science, Artificial Intelligence, and Ethics: Navigating Challenges in the Face of Explosive Growth. In *Annual Review of Biomedical Data Science* (Vol. 7, Issue 1). <https://doi.org/10.1146/annurev-biodatasci-102623-104553>
  83. Hirsbrunner, S. D., Tebbe, M., & Müller-Birn, C. (2024). From critical technical practice to reflexive data science. *Convergence*, 30(1). <https://doi.org/10.1177/13548565221132243>
  84. Misawa, T., Koizumi, A., Tamura, R., & Yoshimi, K. (2025). Exploring utilization of generative AI for research and education in data-driven materials science. *Science and Technology of Advanced Materials: Methods*, 5(1). <https://doi.org/10.1080/27660400.2025.2535956>
  85. Rana, C., & Pardeshi, Prof. Dr. R. K. (2022). A Research Review: Ai and Data Science Applications in the Telecom Industry. *IBMRD's Journal of Management & Research*.  
<https://doi.org/10.17697/ibmrd/2022/v11i2/172620>
  86. Sarker, I. H., Hoque, M. M., Uddin, M. K., & Alsanoosy, T. (2021). Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modeling and Research Directions. *Mobile Networks and Applications*, 26(1). <https://doi.org/10.1007/s11036-020-01650-z>

87. Tawil, A. R. H., Mohamed, M., Schmoor, X., Vlachos, K., & Haidar, D. (2024). Trends and Challenges towards Effective Data-Driven Decision Making in UK Small and Medium-Sized Enterprises: Case Studies and Lessons Learnt from the Analysis of 85 Small and Medium-Sized Enterprises. *Big Data and Cognitive Computing*, 8(7). <https://doi.org/10.3390/bdcc8070079>
88. Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C. W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., ... Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. In *Innovation* (Vol. 2, Issue 4). <https://doi.org/10.1016/j.xinn.2021.100179>
89. Almeida, V., & Nas, E. (2024). Desafios da IA responsável na pesquisa científica. *Revista USP*, 141. <https://doi.org/10.11606/issn.2316-9036.i141p17-28>



De esta edición de *“Metodología de la investigación científica aplicada a la inteligencia artificial y la ciencia de datos: Enfoque general”*, se terminó de editar en la ciudad de Colonia del Sacramento en la República Oriental del Uruguay el 05 de diciembre de 2025

METODOLOGÍA DE LA  
INVESTIGACIÓN CIENTÍFICA  
APLICADA A LA  
INTELIGENCIA ARTIFICIAL Y  
LA CIENCIA DE DATOS:  
**ENFOQUE GENERAL**



ISBN: 978-9915-698-54-0



[www.editorialmarcaribe.es](http://www.editorialmarcaribe.es)