RCMOS – Revista Científica Multidisciplinar O Saber. ISSN: 2675-9128. São Paulo - SP.
Ano VII, v.1 2026 | submissão: 31/07/2026 | aceito: 03/08/2026 | publicação: 06/08/2026

Data Discovery como pilar estratégico para programas de gobernanza de inteligencia artificial: un análisis histórico-evolutivo

Data Discovery como pilar estratégico para programas de governança de inteligência artificial: uma análise histórico-evolutiva

Data Discovery as a strategic pillar for artificial intelligence governance programs: a historical-evolutionary analysis

Luciano da Silva Ferreira

Resumen: Este artículo examina, desde una perspectiva histórico-evolutiva, el papel del Data Discovery como pilar estratégico de los programas de gobernanza de Inteligencia Artificial (IA). Se parte de la premisa de que la madurez de cualquier sistema de gobernanza algorítmica depende directamente de la capacidad organizacional para localizar, clasificar, inventariar y comprender la procedencia de los datos que alimentan los modelos de aprendizaje automático. El estudio recorre la trayectoria que va desde la gestión tradicional de datos corporativos, consolidada a partir de la década de 1980 con los primeros data warehouses, hasta los marcos contemporáneos de gobernanza de IA, como el AI Risk Management Framework del NIST, la norma ISO/IEC 42001:2023 y el Reglamento Europeo de Inteligencia Artificial (AI Act). Se utiliza una metodología de revisión narrativa de la literatura, con análisis documental de normas técnicas, hitos regulatorios y publicaciones científicas indexadas, articulando conceptos de descubrimiento de datos, linaje de datos (data lineage), catalogación automatizada y metadatos de aprendizaje automático. Los resultados indican que el Data Discovery dejó de ser una actividad operativa de TI para convertirse en un requisito estructurante de cumplimiento, explicabilidad y rendición de cuentas (accountability) en sistemas de IA, funcionando como condición previa para la auditabilidad, el mapeo de riesgos y la mitigación de sesgos algorítmicos. Se concluye que las organizaciones que no invierten en capacidades robustas de descubrimiento y catalogación de datos enfrentan limitaciones estructurales para operacionalizar principios éticos y regulatorios de IA, lo que refuerza la tesis de que la gobernanza de datos constituye un fundamento indisociable de la gobernanza algorítmica.

Palabras clave: Data Discovery. Gobernanza de Datos. Gobernanza de Inteligencia Artificial. Linaje de Datos. Cumplimiento Regulatorio.

Resumo: Este artigo examina, sob uma perspectiva histórico-evolutiva, o papel do Data Discovery como pilar estratégico dos programas de governança de Inteligência Artificial (IA). Parte-se da premissa de que a maturidade de qualquer sistema de governança algorítmica depende diretamente da capacidade organizacional de localizar, classificar, inventariar e compreender a proveniência dos dados que alimentam modelos de aprendizado de máquina. O estudo percorre a trajetória que vai da gestão tradicional de dados corporativos, consolidada a partir da década de 1980 com os primeiros data warehouses, até os frameworks contemporâneos de governança de IA, como o AI Risk Management Framework do NIST, a norma ISO/IEC 42001:2023 e o Regulamento Europeu de Inteligência Artificial (AI Act). Utiliza-se metodologia de revisão narrativa da literatura, com análise documental de normas técnicas, marcos regulatórios e publicações científicas indexadas, articulando conceitos de descoberta de dados, linhagem de dados (data lineage), catalogação automatizada e metadados de aprendizado de máquina. Os resultados indicam que o Data Discovery deixou de ser uma atividade operacional de TI para se converter em requisito estruturante de conformidade, explicabilidade e responsabilização (accountability) em sistemas de IA, funcionando como pré-condição para auditabilidade, mapeamento de riscos e mitigação de vieses algorítmicos. Conclui-se que organizações que não investem em capacidades robustas de descoberta e catalogação de dados

enfrentam limitações estruturais para operacionalizar princípios éticos e regulatórios de IA, reforçando a tese de que a governança de dados constitui alicerce indissociável da governança algorítmica.

Palavras-chave: Data Discovery. Governança de Dados. Governança de Inteligência Artificial. Linhagem de Dados. Conformidade Regulatória.

Abstract: This article examines, from a historical-evolutionary perspective, the role of Data Discovery as a strategic pillar of Artificial Intelligence (AI) governance programs. It starts from the premise that the maturity of any algorithmic governance system depends directly on the organizational capacity to locate, classify, inventory, and understand the provenance of the data feeding machine learning models. The study traces the trajectory from traditional corporate data management, consolidated since the 1980s with the first data warehouses, to contemporary AI governance frameworks such as the NIST AI Risk Management Framework, ISO/IEC 42001:2023, and the European Artificial Intelligence Act (AI Act). A narrative literature review methodology is employed, with documentary analysis of technical standards, regulatory milestones, and indexed scientific publications, articulating concepts of data discovery, data lineage, automated cataloging, and machine learning metadata. Results indicate that Data Discovery has ceased to be a mere operational IT activity and has become a structuring requirement for compliance, explainability, and accountability in AI systems, serving as a precondition for auditability, risk mapping, and algorithmic bias mitigation. It is concluded that organizations that fail to invest in robust data discovery and cataloging capabilities face structural limitations in operationalizing ethical and regulatory AI principles, reinforcing the thesis that data governance is an inseparable foundation of algorithmic governance.

Keywords: Data Discovery. Data Governance. Artificial Intelligence Governance. Data Lineage. Regulatory Compliance.

1 INTRODUCCIÓN

El ascenso de los sistemas de Inteligencia Artificial (IA) como infraestructura crítica de decisión en sectores como salud, finanzas, justicia, educación y seguridad pública impuso a las organizaciones un desafío antes restringido a los departamentos técnicos de tecnología de la información: saber con precisión qué datos existen, dónde están almacenados, cuál es su origen, su calidad y su adecuación legal para el uso en modelos algorítmicos. Este desafío, denominado Data Discovery o descubrimiento de datos en la literatura técnica, dejó de ser una actividad auxiliar de mapeo de bases y pasó a ocupar una posición central en los programas de gobernanza de IA implementados por empresas, gobiernos y organismos multilaterales.

El presente artículo parte de la hipótesis de que no es posible construir una gobernanza de IA madura, auditable y éticamente responsable sin que exista previamente una capacidad robusta de descubrimiento, clasificación y seguimiento de datos. Esta hipótesis se sustenta en la constatación histórica de que los primeros grandes escándalos y fallos de sistemas algorítmicos—desde sesgos discriminatorios en sistemas de concesión de crédito hasta fallos

de reconocimiento facial con diferenciación racial—tuvieron como causa raíz recurrente la falta de visibilidad sobre la composición, la procedencia y las limitaciones de los conjuntos de datos utilizados para entrenar los modelos.

Desde esta perspectiva, el objetivo general de este trabajo es analizar, mediante un enfoque histórico-evolutivo, cómo el concepto y la práctica del Data Discovery se desarrollaron desde la gestión tradicional de datos corporativos hasta consolidarse como pilar estratégico de los marcos contemporáneos de gobernanza de IA. Como objetivos específicos, se busca: a) reconstruir la trayectoria histórica de la gobernanza de datos y su relación con el surgimiento de la gobernanza de IA; b) caracterizar técnicamente el Data Discovery y sus principales metodologías y herramientas; c) discutir cómo el descubrimiento de datos se articula con exigencias regulatorias como la Ley General de Protección de Datos brasileña (LGPD), el Reglamento General de Protección de Datos europeo (GDPR) y el Reglamento Europeo de Inteligencia Artificial (AI Act); y d) señalar desafíos y perspectivas futuras para la integración entre la gobernanza de datos y la gobernanza algorítmica.

Metodológicamente, el estudio se caracteriza como una investigación cualitativa, de naturaleza exploratoria y descriptiva, realizada mediante una revisión narrativa de la literatura y análisis documental. Se examinaron artículos científicos indexados en bases como Scopus, Web of Science y Google Scholar, normas técnicas internacionales (como la familia ISO/IEC 27000 y la norma ISO/IEC 42001:2023), marcos institucionales (NIST AI Risk Management Framework y OECD AI Principles) e instrumentos regulatorios (LGPD, GDPR y AI Act), lo que permitió reconstruir una línea de tiempo evolutiva que articula la gobernanza de datos y la gobernanza de IA como fenómenos histórica y técnicamente interdependientes.

El artículo está estructurado en cinco secciones, además de esta introducción. La segunda sección presenta la fundamentación teórica, abordando la génesis de la gobernanza de datos, la trayectoria histórica del Data Discovery y los hitos regulatorios de la gobernanza de IA. La tercera sección analiza específicamente el papel del Data Discovery como pilar estratégico de la gobernanza de IA, detallando la cadena de procedencia de datos, los requisitos de cumplimiento regulatorio y las tecnologías contemporáneas de descubrimiento y catalogación. La cuarta sección discute desafíos, limitaciones y perspectivas futuras. Por último, la quinta sección presenta las consideraciones finales del estudio.

2 FUNDAMENTACIÓN TEÓRICA

2.1 Gobernanza de Datos: Génesis y Evolución Conceptual

La gobernanza de datos, como disciplina formal, tiene sus raíces en la evolución de los sistemas de gestión de bases de datos (SGBD) relacionales, popularizados a partir de la década de 1970 con los trabajos de Edgar F. Codd sobre el modelo relacional. Sin embargo, fue solo a partir de la década de 1990, con la proliferación de data warehouses corporativos y la necesidad de integrar datos procedentes de múltiples sistemas transaccionales, cuando surgió la preocupación sistemática por la calidad, la estandarización y la rendición de cuentas sobre los datos corporativos. Este período corresponde a lo que la literatura de gestión de la información denomina convencionalmente primera generación de la gobernanza de datos, centrada principalmente en controles técnicos de calidad e integridad referencial.

La segunda generación de la gobernanza de datos, situada aproximadamente entre finales de la década de 1990 y la primera década de los años 2000, está marcada por la formalización de marcos metodológicos como DAMA-DMBOK (Data Management Body of Knowledge), organizado por la Data Management Association International. Este marco consolidó diez áreas de conocimiento—entre ellas arquitectura de datos, modelado, calidad, seguridad, metadatos y, de forma central para este estudio, descubrimiento y catálogo de datos —y estableció un vocabulario común y prácticas de referencia para profesionales y organizaciones. Durante este período, la gobernanza de datos aún permanecía mayoritariamente circunscrita a los departamentos de TI y a las áreas de inteligencia de negocios, sin conexión directa con preocupaciones éticas o regulatorias más amplias.

La tercera generación, iniciada alrededor de 2010, fue impulsada por dos fenómenos simultáneos: la explosión del big data—caracterizado por los atributos de volumen, velocidad y variedad, según la formulación seminal de Doug Laney—y la promulgación de marcos regulatorios de protección de datos personales, especialmente el GDPR europeo, aprobado en 2016 y vigente desde 2018, y posteriormente la LGPD brasileña, sancionada en 2018 y vigente desde 2020. Estos marcos regulatorios trasladaron la gobernanza de datos de una preocupación eminentemente técnica a una cuestión de cumplimiento legal, derechos fundamentales y responsabilidad corporativa, exigiendo que las organizaciones pudieran demostrar en cualquier momento qué datos personales poseían, dónde estaban almacenados, con qué finalidad habían sido recopilados y durante cuánto tiempo serían conservados.

Es precisamente en este contexto de exigencia regulatoria donde el Data Discovery se consolida como disciplina técnica autónoma, pues la respuesta a solicitudes de titulares de datos, auditorías regulatorias e incidentes de seguridad hizo inviable depender exclusivamente

de inventarios manuales de datos, históricamente desactualizados e incompletos en las grandes organizaciones.

2.2 Data Discovery: Trayectoria Histórica y Evolución Tecnológica

El término Data Discovery se utiliza en la literatura técnica con dos sentidos complementarios que conviene distinguir. En el campo del análisis de datos y la inteligencia de negocios, popularizado por herramientas como Tableau y Qlik durante los años 2000, Data Discovery designa el proceso interactivo y exploratorio de identificación de patrones, tendencias y correlaciones en conjuntos de datos, realizado normalmente por analistas de negocio sin depender de equipos especializados de TI. En el campo de la gobernanza de datos y la seguridad de la información—sentido que orienta centralmente este artículo—Data Discovery se refiere al proceso sistemático y cada vez más automatizado de localizar, identificar, clasificar y catalogar datos dispersos por sistemas, aplicaciones, bases de datos estructuradas y no estructuradas, entornos de nube y repositorios de archivos, con el fin de construir un inventario confiable sobre la existencia, la naturaleza y la sensibilidad de esos datos.

Históricamente, la primera generación de herramientas de descubrimiento de datos, durante las décadas de 1990 y 2000, operaba mediante exploraciones basadas en expresiones regulares y diccionarios de patrones, capaces de identificar formatos estructurados como números de documentos de identidad, tarjetas de crédito y direcciones de correo electrónico en bases de datos relacionales. Este enfoque, todavía utilizado como capa base en diversas soluciones comerciales, presentaba limitaciones importantes ante datos no estructurados, como documentos de texto libre, imágenes, audios y videos, que pasaron a representar una proporción creciente del volumen de datos corporativos a partir de la década de 2010.

La segunda generación de herramientas, consolidada entre 2012 y 2018, incorporó técnicas de procesamiento de lenguaje natural (PLN) y aprendizaje automático supervisado para la clasificación semántica de datos, lo que permitió identificar no solo patrones sintácticos, sino también el significado contextual de campos y documentos—por ejemplo, distinguir un número de nueve dígitos que representa un documento de identificación fiscal de un número telefónico con formato similar. Este avance tecnológico coincidió con la maduración de plataformas de big data, como el ecosistema Hadoop y, posteriormente, las arquitecturas de data lake, que agregaban volúmenes masivos de datos heterogéneos sin un esquema

previamente definido (enfoque schema-on-read), haciendo del descubrimiento automatizado una necesidad operativa inaplazable.

La tercera y actual generación de Data Discovery, en consolidación desde aproximadamente 2019, se caracteriza por la integración entre catalogación automatizada de metadatos, linaje de datos (data lineage) de extremo a extremo y, más recientemente, seguimiento específico de metadatos de aprendizaje automático—como conjuntos de datos de entrenamiento, versiones de modelos, hiperparámetros y métricas de rendimiento. Esta generación es impulsada tanto por la adopción de arquitecturas de datos distribuidas en la nube como por la necesidad, impuesta por los marcos de gobernanza de IA, de reconstruir de forma auditable toda la cadena de procedencia de un dato, desde su origen hasta su utilización en decisiones algorítmicas automatizadas.

2.3 Gobernanza de Inteligencia Artificial: Hitos Regulatorios e Institucionales

La gobernanza de Inteligencia Artificial, entendida como el conjunto de estructuras, políticas, procesos y controles orientados a garantizar el desarrollo y el uso responsable, seguro y ético de los sistemas de IA, es un fenómeno considerablemente más reciente que la gobernanza de datos y se consolidó institucionalmente sobre todo a partir de 2019. Ese año, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) publicó sus Principios de IA, adoptados por decenas de países miembros y no miembros, estableciendo directrices de transparencia, robustez, seguridad y rendición de cuentas que influirían prácticamente en todos los marcos nacionales posteriores.

En 2021, la Comisión Europea presentó la propuesta del Reglamento Europeo de Inteligencia Artificial, denominado AI Act, aprobado formalmente en 2024 y con entrada en vigor escalonada durante 2025 y 2026. El AI Act adopta un enfoque basado en riesgos, clasificando los sistemas de IA en categorías de riesgo inaceptable, alto riesgo, riesgo limitado y riesgo mínimo, e impone, para los sistemas de alto riesgo, obligaciones expresas de documentación técnica, gobernanza de los datos de entrenamiento, validación y prueba, y trazabilidad (registro automático), lo que evidencia una conexión normativa directa entre la gobernanza de datos y la gobernanza algorítmica.

Paralelamente, el Instituto Nacional de Estándares y Tecnología de los Estados Unidos (NIST) publicó en 2023 el AI Risk Management Framework (AI RMF), estructurado en cuatro funciones centrales—Gobernar, Mapear, Medir y Gestionar—que exige explícitamente a las organizaciones la capacidad de mapear la procedencia, la calidad y la representatividad de los

datos utilizados en sistemas de IA como condición previa para la gestión de riesgos algorítmicos. En el mismo período, la Organización Internacional de Normalización (ISO), junto con la Comisión Electrotécnica Internacional (IEC), publicó la norma ISO/IEC 42001:2023, primera norma certificable de sistemas de gestión de IA, que dedica requisitos específicos a la calidad y la procedencia de los datos utilizados en el ciclo de vida de los sistemas de IA.

En Brasil, el debate regulatorio sobre IA avanzó mediante el Proyecto de Ley n.º 2338/2023, en tramitación en el Congreso Nacional, que propone un marco legal para el uso de la Inteligencia Artificial inspirado en gran medida en el enfoque basado en riesgos del AI Act europeo, así como mediante la actuación de la Autoridad Nacional de Protección de Datos (ANPD), que ya regula, en el ámbito de la LGPD, el tratamiento automatizado de datos personales con fines de toma de decisiones, incluido el derecho a la revisión de decisiones automatizadas.

La convergencia entre estos marcos regulatorios revela un patrón histórico consistente: ningún marco de gobernanza de IA, ya sea de naturaleza regulatoria o técnica y voluntaria, prescinde de mecanismos robustos de conocimiento sobre los datos subyacentes a los sistemas algorítmicos. Es precisamente esta convergencia la que fundamenta la tesis central de este artículo, desarrollada en la sección siguiente.

3 DATA DISCOVERY COMO PILAR ESTRATÉGICO DE LA GOBERNANZA DE INTELIGENCIA ARTIFICIAL

3.1 La Cadena de Procedencia: del Dato Bruto a la Decisión Algorítmica

Todo sistema de Inteligencia Artificial, independientemente de su arquitectura técnica, resulta de una cadena de transformaciones que se inicia con la recopilación o generación de datos brutos y culmina en una inferencia, predicción o decisión automatizada. Esta cadena—que la literatura técnica denomina cadena de procedencia de datos (data provenance chain)—comprende normalmente las etapas de recopilación, almacenamiento, integración, limpieza, etiquetado, selección de muestras de entrenamiento, ingeniería de características, entrenamiento del modelo, validación, implantación (deployment) y monitoreo continuo. En cada una de estas etapas, decisiones técnicas aparentemente neutrales pueden introducir distorsiones, sesgos o pérdidas de calidad que se propagan y amplifican a lo largo de la cadena.

En este contexto, el Data Discovery actúa como una capa de instrumentación que permite reconstruir esta cadena de forma auditable. Sin mecanismos de descubrimiento y catalogación de datos, una organización es incapaz de responder con precisión y rapidez a preguntas centrales para la gobernanza de IA, tales como: qué conjuntos de datos alimentaron un modelo determinado; si esos conjuntos contienen datos personales sensibles; si existe una representatividad demográfica adecuada en las muestras de entrenamiento; si hubo consentimiento o base legal para el uso de un dato determinado; y si es posible eliminar o corregir un dato específico ante una solicitud del titular sin comprometer la integridad del modelo ya entrenado.

Esta incapacidad de respuesta no es meramente teórica. Estudios de caso ampliamente discutidos en la literatura sobre ética en IA—como el sistema COMPAS de evaluación del riesgo de reincidencia penal utilizado en tribunales estadounidenses y sistemas comerciales de reconocimiento facial con un rendimiento significativamente inferior para tonos de piel más oscuros—tienen en común que, en el momento del desarrollo, los desarrolladores no contaban con visibilidad suficiente sobre la composición demográfica y la representatividad de los datos de entrenamiento utilizados. Esta brecha de visibilidad es, en esencia, un fallo de Data Discovery: la ausencia de un inventario confiable que permitiera identificar previamente el desequilibrio muestral que más tarde se manifestaría como discriminación algorítmica en producción.

3.2 Descubrimiento de Datos y Cumplimiento Regulatorio

La relación entre Data Discovery y cumplimiento regulatorio puede observarse en tres frentes normativos centrales. En primer lugar, la LGPD brasileña y el GDPR europeo exigen que las organizaciones mantengan registros de las actividades de tratamiento de datos personales (Records of Processing Activities, en el vocabulario del GDPR), lo que presupone un conocimiento actualizado de qué datos personales existen, dónde están almacenados y con qué finalidad se tratan. En entornos corporativos con múltiples sistemas heredados y plataformas en la nube, este conocimiento solo es viable a escala mediante herramientas automatizadas de descubrimiento y clasificación de datos.

En segundo lugar, el AI Act europeo establece, para los sistemas de alto riesgo, la obligación de que los conjuntos de datos de entrenamiento, validación y prueba sean pertinentes, representativos, estén libres de errores en la medida de lo posible y sean completos, teniendo en cuenta las características específicas del contexto de uso previsto. El cumplimiento

de esta obligación depende directamente de mecanismos de descubrimiento y perfilado de datos (data profiling) capaces de medir, de forma cuantitativa y documentada, la integridad, la representatividad y la calidad de los datos de entrenamiento antes y durante el ciclo de vida del modelo.

En tercer lugar, el AI RMF del NIST y la norma ISO/IEC 42001:2023 convergen al exigir que las organizaciones mantengan documentación técnica detallada sobre el origen, el procesamiento y las limitaciones de los datos utilizados en sistemas de IA, materializada con frecuencia en artefactos como datasheets for datasets—modelo de documentación estructurada de conjuntos de datos propuesto en la literatura académica e incorporado progresivamente a las prácticas corporativas—y model cards, que documentan el rendimiento y las limitaciones de los modelos de aprendizaje automático en diferentes subgrupos poblacionales. La elaboración de estos artefactos es, en la práctica, imposible sin procesos maduros de Data Discovery, ya que exige un conocimiento preciso de la procedencia, la composición y las características estadísticas de los datos subyacentes.

3.3 Tecnologías y Métodos Contemporáneos de Data Discovery Aplicados a la IA

Las soluciones tecnológicas contemporáneas de Data Discovery orientadas a la gobernanza de IA articulan normalmente cuatro capacidades técnicas centrales. La primera es la catalogación automatizada de datos (data cataloging), que consiste en la exploración sistemática de fuentes de datos estructuradas y no estructuradas para construir un repositorio centralizado de metadatos, incluidos esquemas, tipos de datos, volumen, propietarios responsables y clasificaciones de sensibilidad. Las plataformas contemporáneas de catalogación incorporan algoritmos de aprendizaje automático para inferir automáticamente relaciones semánticas entre campos de datos dispersos en diferentes sistemas, reduciendo la dependencia de documentación manual históricamente incompleta.

La segunda capacidad es la clasificación de datos sensibles (data classification), que aplica técnicas de reconocimiento de patrones, expresiones regulares, diccionarios semánticos y, más recientemente, modelos de lenguaje para identificar categorías de datos personales y sensibles de acuerdo con taxonomías regulatorias específicas—como datos de salud, origen racial o étnico, orientación sexual y datos biométricos—, categorías que reciben un tratamiento jurídico diferenciado tanto en la LGPD como en el GDPR y el AI Act.

La tercera capacidad es el linaje de datos (data lineage), que consiste en el seguimiento automatizado del flujo de datos desde su origen hasta sus puntos de consumo, incluidas todas

las transformaciones intermedias. Las herramientas contemporáneas de linaje funcionan mediante la instrumentación de pipelines de datos (análisis de código SQL, scripts de ingeniería de datos y orquestadores de flujos de trabajo), lo que permite reconstruir de forma automática y visual el trayecto recorrido por un dato específico, capacidad esencial para investigaciones de incidentes, auditorías regulatorias y análisis de causa raíz de sesgos algorítmicos.

La cuarta capacidad, de consolidación más reciente y específica del contexto de IA, es el seguimiento de metadatos de aprendizaje automático (ML metadata tracking), operacionalizado mediante plataformas de MLOps que registran de forma versionada qué conjuntos de datos se utilizaron en cada experimento de entrenamiento, qué transformaciones se aplicaron, qué hiperparámetros se utilizaron y qué métricas de rendimiento—incluidas métricas de equidad algorítmica desglosadas por subgrupo poblacional—se obtuvieron. Esta capacidad extiende el alcance tradicional del Data Discovery, originalmente circunscrito a los datos almacenados en repositorios, al propio ciclo de vida de desarrollo de modelos de IA, en una convergencia técnica que la literatura reciente denomina gobernanza de datos y modelos (data-and-model governance).

3.4 Línea de Tiempo Evolutiva: de la Gobernanza de Datos a la Gobernanza de Inteligencia Artificial

La síntesis histórica desarrollada a lo largo de esta sección permite proponer una periodización evolutiva que articula gobernanza de datos, Data Discovery y gobernanza de IA, organizada en cinco fases sucesivas y acumulativas. La primera fase, situada entre las décadas de 1970 y 1980, corresponde a la consolidación de los sistemas de gestión de bases de datos relacionales y al surgimiento de controles rudimentarios de calidad de datos, sin preocupación sistemática por el descubrimiento o la catalogación.

La segunda fase, entre 1990 y 2005, corresponde a la formalización de la gobernanza de datos como disciplina de gestión, con la consolidación de data warehouses, marcos como DAMA-DMBOK y las primeras herramientas de descubrimiento basadas en patrones sintácticos, todavía restringidas a datos estructurados y a finalidades predominantemente analíticas de inteligencia de negocios.

La tercera fase, entre 2005 y 2016, está marcada por la explosión del big data, la adopción de arquitecturas de data lake y la incorporación del procesamiento de lenguaje natural y el aprendizaje automático a las herramientas de descubrimiento, ampliando su capacidad para operar sobre datos no estructurados a gran escala.

La cuarta fase, entre 2016 y 2023, corresponde a la consolidación regulatoria de la protección de datos personales, con la aprobación del GDPR y la LGPD, que transformaron el Data Discovery de una conveniencia analítica en una obligación de cumplimiento legal. Al mismo tiempo, con los Principios de IA de la OCDE de 2019, comenzó la institucionalización de la gobernanza de IA como campo autónomo.

La quinta y actual fase, iniciada en 2023 y aún en consolidación, se caracteriza por la convergencia explícita entre la gobernanza de datos y la gobernanza de IA, materializada en la ISO/IEC 42001:2023, el AI RMF del NIST y el AI Act europeo, que exigen capacidades avanzadas de descubrimiento, catalogación, linaje y seguimiento de metadatos como condición técnica previa para el cumplimiento regulatorio y la gestión de riesgos algorítmicos. Esta periodización muestra que el Data Discovery no surgió como respuesta a la gobernanza de IA, sino como una disciplina previamente consolidada que se reconfiguró y adquirió una centralidad estratégica ante las necesidades específicas impuestas por los sistemas algorítmicos contemporáneos.

4 DISCUSIÓN: DESAFÍOS, LIMITACIONES Y PERSPECTIVAS FUTURAS

A pesar de los avances tecnológicos y regulatorios descritos, la integración entre Data Discovery y gobernanza de IA enfrenta importantes desafíos estructurales. El primero se refiere a la escala y la heterogeneidad de los entornos de datos contemporáneos, caracterizados por la coexistencia de sistemas heredados, plataformas en la nube de múltiples proveedores, aplicaciones de software como servicio (SaaS) y fuentes de datos de terceros, lo que convierte la construcción de un inventario verdaderamente completo y actualizado en una meta asintótica, raramente alcanzada en su totalidad incluso por las organizaciones más maduras.

El segundo desafío se refiere a la limitación técnica de las propias herramientas de descubrimiento automatizado ante datos no estructurados de naturaleza multimodal—como combinaciones de texto, imagen, audio y video utilizadas en el entrenamiento de modelos fundacionales contemporáneos—, para los cuales los métodos de clasificación semántica todavía presentan tasas de error superiores a las observadas en datos estructurados tradicionales, lo que introduce incertidumbre residual incluso en programas de gobernanza técnicamente avanzados.

El tercer desafío es de naturaleza organizacional y cultural, y quizá el más persistente: la gobernanza de datos y la gobernanza de IA suelen residir en estructuras organizacionales distintas—la primera en áreas de tecnología de la información y protección de datos, y la

segunda en comités de ética, áreas jurídicas o unidades de ciencia de datos—, lo que genera brechas de comunicación, duplicación de esfuerzos y, en casos extremos, inventarios de datos e inventarios de modelos de IA que no se comunican entre sí, comprometiendo la trazabilidad de extremo a extremo exigida por los marcos regulatorios contemporáneos.

El cuarto desafío se refiere a la tensión entre la granularidad exigida por la auditabilidad regulatoria y los costos computacionales y financieros de mantener el linaje de datos y los metadatos de aprendizaje automático a escala de producción, especialmente en organizaciones que operan cientos o miles de modelos simultáneamente, un escenario cada vez más común ante la proliferación de aplicaciones basadas en modelos de lenguaje de gran tamaño (LLM) y arquitecturas de IA generativa.

Como perspectivas futuras, se observa una tendencia de convergencia técnica entre plataformas de gobernanza de datos y plataformas de MLOps, con el surgimiento de soluciones integradas que tratan el dato y el modelo como artefactos de una misma cadena de procedencia unificada, enfoque que la literatura reciente denomina gobernanza unificada de datos e IA. Asimismo, existe un interés creciente en técnicas de descubrimiento y clasificación basadas en modelos de lenguaje de gran tamaño, que prometen mayor precisión semántica en la identificación de datos sensibles en contextos no estructurados, aunque introducen nuevos desafíos relacionados con la explicabilidad y la confiabilidad de las propias herramientas de descubrimiento. Finalmente, es razonable prever que futuros marcos regulatorios, en Brasil y otras jurisdicciones, tenderán a explicitar con creciente detalle requisitos técnicos mínimos de descubrimiento, catalogación y linaje de datos como condición de cumplimiento para sistemas de IA de alto riesgo, consolidando definitivamente el Data Discovery como infraestructura crítica—y ya no accesoria—de la gobernanza algorítmica.

5 CONSIDERACIONES FINALES

Este artículo buscó demostrar, mediante un análisis histórico-evolutivo, que el Data Discovery constituye un pilar estratégico indisociable de los programas contemporáneos de gobernanza de Inteligencia Artificial. La reconstrucción de la trayectoria que va desde la consolidación de los sistemas de gestión de bases de datos relacionales en las décadas de 1970 y 1980 hasta la convergencia regulatoria materializada en el AI Act europeo, el AI Risk Management Framework del NIST y la norma ISO/IEC 42001:2023 revela un patrón histórico consistente: la capacidad de localizar, clasificar y rastrear la procedencia de los datos siempre

precedió, técnicamente, a la capacidad de gobernar los sistemas construidos a partir de esos datos.

La principal contribución de este estudio reside en la sistematización de una periodización evolutiva en cinco fases, articulando conceptos hasta entonces tratados de forma relativamente fragmentada en la literatura—gobernanza de datos, Data Discovery y gobernanza de IA—y evidenciando su interdependencia técnica y regulatoria. Con base en esta periodización, se argumenta que las organizaciones que no invierten en capacidades maduras de descubrimiento, catalogación y linaje de datos enfrentan limitaciones estructurales, y no meramente operativas, para poner en práctica principios éticos y cumplir exigencias regulatorias de gobernanza de IA, puesto que el cumplimiento de requisitos de representatividad, calidad, explicabilidad y auditabilidad de sistemas algorítmicos presupone un conocimiento previo y confiable de los datos subyacentes.

Como limitación de este estudio, destaca su carácter eminentemente teórico y documental, sin validación empírica directa mediante estudios de caso organizacionales profundos o investigación de campo. Como agenda de futuras investigaciones, se propone realizar estudios empíricos que midan, en organizaciones de diferentes sectores y tamaños, la correlación entre el grado de madurez de Data Discovery y los indicadores efectivos de cumplimiento y mitigación de riesgos en programas de gobernanza de IA, así como investigar en profundidad modelos de gobernanza unificada de datos y modelos de aprendizaje automático, campo aún incipiente y con elevado potencial de desarrollo académico y práctico en los próximos años.

REFERENCIAS

BRASIL. Ley n.º 13.709, de 14 de agosto de 2018. Ley General de Protección de Datos Personales (LGPD). Brasília: Diario Oficial de la Unión, 2018.

BRASIL. Proyecto de Ley n.º 2338, de 2023. Dispone sobre el uso de la Inteligencia Artificial. Brasília: Senado Federal, 2023.

COMISIÓN EUROPEA. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial). Bruselas: Diario Oficial de la Unión Europea, 2024.

DAMA INTERNATIONAL. DAMA-DMBOK: Cuerpo de Conocimiento de Gestión de Datos. 2.ª ed. Denville: Technics Publications, 2017.

FLORIDI, L.; COWLS, J. Un marco unificado de cinco principios para la IA en la sociedad. Harvard Data Science Review, Cambridge, v. 1, n. 1, 2019.

GEBRU, T. et al. Fichas técnicas para conjuntos de datos. Communications of the ACM, Nueva York, v. 64, n. 12, p. 86-92, 2021.

INSTITUTO NACIONAL DE ESTÁNDARES Y TECNOLOGÍA. Marco de Gestión de Riesgos de IA (AI RMF 1.0). Gaithersburg: NIST, 2023.

JOBIN, A.; IENCA, M.; VAYENA, E. El panorama mundial de las directrices éticas para la IA. Nature Machine Intelligence, Londres, v. 1, n. 9, p. 389-399, 2019.

KITCHIN, R. La revolución de los datos: big data, datos abiertos, infraestructuras de datos y sus consecuencias. Londres: Sage Publications, 2014.

LANEY, D. Gestión de datos 3D: control del volumen, la velocidad y la variedad de los datos. Stamford: META Group Research Note, 2001.

MITCHELL, M. et al. Tarjetas de modelo para la presentación de informes sobre modelos. En: CONFERENCIA SOBRE EQUIDAD, RENDICIÓN DE CUENTAS Y TRANSPARENCIA, 2019, Atlanta. Actas [...]. Nueva York: ACM, 2019. p. 220-229.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. Recomendación del Consejo sobre Inteligencia Artificial, OECD/LEGAL/0449. París: OECD, 2019.

ORGANIZACIÓN INTERNACIONAL DE NORMALIZACIÓN. ISO/IEC 27001:2022—Seguridad de la información, ciberseguridad y protección de la privacidad—Sistemas de gestión de la seguridad de la información—Requisitos. Ginebra: ISO, 2022.

ORGANIZACIÓN INTERNACIONAL DE NORMALIZACIÓN. ISO/IEC 42001:2023—Tecnología de la información—Inteligencia artificial—Sistema de gestión. Ginebra: ISO, 2023.

PARLAMENTO EUROPEO; CONSEJO DE LA UNIÓN EUROPEA. Reglamento (UE) 2016/679, de 27 de abril de 2016 (Reglamento General de Protección de Datos—GDPR). Bruselas: Diario Oficial de la Unión Europea, 2016.

RUSSELL, S.; NORVIG, P. Inteligencia Artificial: un enfoque moderno. 4.ª ed. Rio de Janeiro: GEN LTC, 2022.

ZUBOFF, S. La era del capitalismo de vigilancia. Rio de Janeiro: Intrínseca, 2021.