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 governança de inteligência artificial: uma análise histórico-evolutiva

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

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

Luciano da Silva Ferreira

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.

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.

1 INTRODUÇÃO

A ascensão dos sistemas de Inteligência Artificial (IA) como infraestrutura crítica de decisão em setores como saúde, finanças, justiça, educação e segurança pública impôs às organizações um desafio antes restrito a departamentos técnicos de tecnologia da informação: saber, com precisão, quais dados existem, onde estão armazenados, qual sua origem, sua qualidade e sua adequação legal para uso em modelos algorítmicos. Esse desafio, denominado na literatura técnica como Data Discovery, ou descoberta de dados, deixou de ser uma atividade acessória de mapeamento de bases e passou a ocupar posição central nos programas de governança de IA implementados por empresas, governos e organismos multilaterais.

O presente artigo parte da hipótese de que não é possível construir uma governança de IA madura, auditável e eticamente responsável sem que exista, previamente, uma capacidade robusta de descoberta, classificação e rastreamento de dados. Tal hipótese se sustenta na constatação histórica de que os primeiros grandes escândalos e falhas de sistemas algorítmicos — de vieses discriminatórios em sistemas de concessão de crédito a falhas de reconhecimento

facial com recorte racial — tiveram como causa-raiz recorrente a ausência de visibilidade sobre a composição, a procedência e as limitações dos conjuntos de dados utilizados no treinamento dos modelos.

Sob essa perspectiva, o objetivo geral deste trabalho é analisar, por meio de uma abordagem histórico-evolutiva, como o conceito e a prática de Data Discovery se desenvolveram desde a gestão tradicional de dados corporativos até se consolidarem como pilar estratégico dos frameworks contemporâneos de governança de IA. Como objetivos específicos, busca-se: a) reconstituir a trajetória histórica da governança de dados e sua relação com a emergência da governança de IA; b) caracterizar tecnicamente o Data Discovery e suas principais metodologias e ferramentas; c) discutir de que forma a descoberta de dados se articula com exigências regulatórias como a Lei Geral de Proteção de Dados (LGPD), o Regulamento Geral de Proteção de Dados europeu (GDPR) e o Regulamento Europeu de Inteligência Artificial (AI Act); e d) apontar desafios e perspectivas futuras para a integração entre governança de dados e governança algorítmica.

Metodologicamente, o estudo caracteriza-se como uma pesquisa qualitativa, de natureza exploratória e descritiva, operacionalizada por meio de revisão narrativa da literatura e análise documental. Foram examinados artigos científicos indexados em bases como Scopus, Web of Science e Google Scholar, normas técnicas internacionais (a exemplo da família ISO/IEC 27000 e da norma ISO/IEC 42001:2023), frameworks institucionais (NIST AI Risk Management Framework, OECD AI Principles) e marcos regulatórios (LGPD, GDPR e AI Act), permitindo a reconstrução de uma linha do tempo evolutiva que articula governança de dados e governança de IA como fenômenos histórica e tecnicamente interdependentes.

O artigo está estruturado em cinco seções, além desta introdução. A segunda seção apresenta a fundamentação teórica, abordando a gênese da governança de dados, a trajetória histórica do Data Discovery e os marcos regulatórios da governança de IA. A terceira seção discute especificamente o papel do Data Discovery como pilar estratégico da governança de IA detalhando a cadeia de proveniência de dados, os requisitos de conformidade regulatória e as tecnologias contemporâneas de descoberta e catalogação. A quarta seção discute desafios, limitações e perspectivas futuras. Por fim, a quinta seção apresenta as considerações finais do estudo.

2 FUNDAMENTAÇÃO TEÓRICA

Ano VII, v.1 2026 | submissão: 31/07/2026 | aceito: 03/08/2026 | publicação: 06/08/2026

2.1 Governança de Dados: Gênese e Evolução Conceitual

A governança de dados, enquanto disciplina formal, tem suas raízes na evolução dos sistemas de gerenciamento de banco de dados (SGBD) relacionais, popularizados a partir da década de 1970 com os trabalhos de Edgar F. Codd sobre o modelo relacional. Contudo, foi somente a partir da década de 1990, com a proliferação de data warehouses corporativos e a necessidade de integrar dados provenientes de múltiplos sistemas transacionais, que emergiu a preocupação sistemática com qualidade, padronização e responsabilização sobre os dados corporativos. Esse período corresponde ao que a literatura de gestão da informação convencionou chamar de primeira geração da governança de dados, centrada primariamente em controles técnicos de qualidade e integridade referencial.

A segunda geração da governança de dados, situada aproximadamente entre o final dos anos 1990 e a primeira década dos anos 2000, é marcada pela formalização de arcabouços metodológicos como o DAMA-DMBOK (Data Management Body of Knowledge), organizado pela Data Management Association International. Esse arcabouço consolidou dez áreas de conhecimento — entre elas arquitetura de dados, modelagem, qualidade, segurança, metadados e, centralmente para este estudo, a descoberta e o catálogo de dados — estabelecendo vocabulário comum e práticas de referência para profissionais e organizações. Nesse período, a governança de dados ainda permanecia majoritariamente circunscrita a departamentos de TI e áreas de business intelligence, sem conexão direta com preocupações éticas ou regulatórias mais amplas.

A terceira geração, iniciada por volta de 2010, é impulsionada por dois fenômenos concomitantes: a explosão da big data — caracterizado pelos atributos de volume, velocidade e variedade, conforme a formulação seminal de Doug Laney — e a promulgação de marcos regulatórios de proteção de dados pessoais, com destaque para o GDPR europeu, aprovado em 2016 e vigente a partir de 2018, e posteriormente a LGPD brasileira, sancionada em 2018 e vigente desde 2020. Esses marcos regulatórios deslocaram a governança de dados de uma preocupação eminentemente técnica para uma questão de conformidade legal, direitos fundamentais e responsabilização corporativa, exigindo das organizações a capacidade de demonstrar, a qualquer momento, quais dados pessoais possuíam, onde estavam armazenados, com que finalidade haviam sido coletados e por quanto tempo seriam retidos.

É precisamente nesse contexto de exigência regulatória que o Data Discovery se consolida como disciplina técnica autônoma, pois a resposta a solicitações de titulares de dados, a auditorias regulatórias e a incidentes de segurança tornou inviável depender exclusivamente

de inventários manuais de dados, historicamente desatualizados e incompletos nas grandes organizações.

2.2 Data Discovery: Trajetória Histórica e Evolução Tecnológica

O termo Data Discovery é utilizado na literatura técnica com dois sentidos complementares, que convém distinguir. No campo da análise de dados e da inteligência de negócios, popularizado por ferramentas como Tableau e Qlik ao longo dos anos 2000, Data Discovery designa o processo interativo e exploratório de identificação de padrões, tendências e correlações em conjuntos de dados, tipicamente conduzido por analistas de negócio sem dependência de equipes especializadas de TI. Já no campo da governança de dados e da segurança da informação, sentido que orienta centralmente este artigo, Data Discovery refere-se ao processo sistemático e, cada vez mais, automatizado de localizar, identificar, classificar e catalogar dados dispersos por sistemas, aplicações, bancos de dados estruturados e não estruturados, ambientes em nuvem e repositórios de arquivos, de modo a construir um inventário confiável sobre a existência, a natureza e a sensibilidade desses dados.

Historicamente, a primeira geração de ferramentas de descoberta de dados, ao longo dos anos 1990 e 2000, operava por meio de varreduras baseadas em expressões regulares e dicionários de padrões, capazes de identificar formatos estruturados como números de documentos de identificação, cartões de crédito e endereços de correio eletrônico em bancos de dados relacionais. Essa abordagem, ainda hoje utilizada como camada de base em diversas soluções comerciais, apresentava limitações relevantes diante de dados não estruturados, como documentos de texto livre, imagens, áudios e vídeos, que passaram a representar parcela crescente do volume de dados corporativos a partir da década de 2010.

A segunda geração de ferramentas, consolidada entre 2012 e 2018, incorporou técnicas de processamento de linguagem natural (PLN) e aprendizado de máquina supervisionado para classificação semântica de dados, permitindo identificar não apenas padrões sintáticos, mas também o significado contextual de campos e documentos — por exemplo, distinguir um número de nove dígitos que representa um documento de identificação fiscal de um número de telefone com formatação semelhante. Esse avanço tecnológico coincidiu com a maturação de plataformas de big data, como o ecossistema Hadoop e, posteriormente, arquiteturas de data lake, que agregavam volumes massivos de dados heterogêneos sem esquema definido previamente (abordagem schema-on-read), tornando a descoberta automatizada uma necessidade operacional inadiável.

A terceira e atual geração de Data Discovery, em consolidação desde aproximadamente 2019, caracteriza-se pela integração entre catalogação automatizada de metadados, linhagem de dados (data lineage) de ponta a ponta e, mais recentemente, rastreamento específico de metadados de aprendizado de máquina — como conjuntos de dados de treinamento, versões de modelos, hiperparâmetros e métricas de desempenho. Essa geração é impulsionada tanto pela adoção de arquiteturas de dados distribuídas em nuvem quanto pela necessidade, imposta por frameworks de governança de IA, de reconstituir de forma auditável toda a cadeia de proveniência de um dado, desde sua origem até sua utilização em decisões algorítmicas automatizadas.

2.3 Governança de Inteligência Artificial: Marcos Regulatórios e Institucionais

A governança de Inteligência Artificial, compreendida como o conjunto de estruturas, políticas, processos e controles voltados a assegurar o desenvolvimento e o uso responsável, seguro e ético de sistemas de IA, é fenômeno consideravelmente mais recente que a governança de dados, tendo se consolidado institucionalmente sobretudo a partir de 2019. Nesse ano, a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) publicou seus Princípios de IA, adotados por dezenas de países membros e não membros, estabelecendo diretrizes de transparência, robustez, segurança e responsabilização que viriam a influenciar praticamente todos os frameworks nacionais subsequentes.

Em 2021, a Comissão Europeia apresentou a proposta do Regulamento Europeu de Inteligência Artificial, o chamado AI Act, aprovado formalmente em 2024 e com entrada em vigor escalonada ao longo de 2025 e 2026. O AI Act adota abordagem baseada em risco, classificando sistemas de IA em categorias de risco inaceitável, alto risco, risco limitado e risco mínimo, e impõe, para sistemas de alto risco, obrigações expressas de documentação técnica, governança de dados de treinamento, validação e teste, e rastreabilidade (logging automático), evidenciando uma conexão normativa direta entre governança de dados e governança algorítmica.

Paralelamente, o Instituto Nacional de Padrões e Tecnologia dos Estados Unidos (NIST) publicou, em 2023, o AI Risk Management Framework (AI RMF), estruturado em quatro funções centrais — Governar, Mapear, Medir e Gerenciar — que explicitamente demandam das organizações a capacidade de mapear a proveniência, a qualidade e a representatividade dos dados utilizados em sistemas de IA como pré-condição para a gestão de riscos algorítmicos. No mesmo período, a Organização Internacional de Normalização (ISO), em conjunto com a

Comissão Eletrotécnica Internacional (IEC), publicou a norma ISO/IEC 42001:2023, primeira norma certificável de sistema de gestão de IA, que dedica requisitos específicos à qualidade e à proveniência dos dados utilizados no ciclo de vida de sistemas de IA.

No Brasil, o debate regulatório sobre IA avançou por meio do Projeto de Lei nº 2338/2023, em tramitação no Congresso Nacional, que propõe um marco legal para o uso de Inteligência Artificial inspirado, em larga medida, na abordagem baseada em risco do AI Act europeu, bem como pela atuação da Autoridade Nacional de Proteção de Dados (ANPD), que já regula, no âmbito da LGPD, o tratamento automatizado de dados pessoais para fins de tomada de decisão, incluindo o direito à revisão de decisões automatizadas.

A convergência entre esses marcos regulatórios revela um padrão histórico consistente: nenhum framework de governança de IA, seja de natureza regulatória, seja de natureza técnica-voluntária, prescinde de mecanismos robustos de conhecimento sobre os dados subjacentes aos sistemas algorítmicos. É precisamente essa convergência que fundamenta a tese central deste artigo, desenvolvida na seção seguinte.

3 DATA DISCOVERY COMO PILAR ESTRATÉGICO DA GOVERNANÇA DE INTELIGÊNCIA ARTIFICIAL

3.1 A Cadeia de Proveniência: do Dado Bruto à Decisão Algorítmica

Todo sistema de Inteligência Artificial, independentemente de sua arquitetura técnica, resulta de uma cadeia de transformações que se inicia na coleta ou geração de dados brutos e culmina em uma inferência, predição ou decisão automatizada. Essa cadeia — que a literatura técnica denomina cadeia de proveniência de dados (data provenance chain) — envolve, tipicamente, as etapas de coleta, armazenamento, integração, limpeza, rotulagem, seleção de amostras de treinamento, engenharia de atributos, treinamento do modelo, validação, implantação (deployment) e monitoramento contínuo. Em cada uma dessas etapas, decisões técnicas aparentemente neutras podem introduzir distorções, vieses ou perdas de qualidade que se propagam e se amplificam ao longo da cadeia.

O Data Discovery atua, nesse contexto, como camada de instrumentação que possibilita reconstituir essa cadeia de forma auditável. Sem mecanismos de descoberta e catalogação de dados, uma organização é incapaz de responder, com precisão e tempestividade, a perguntas centrais para a governança de IA, tais como: quais conjuntos de dados alimentaram determinado

modelo; se esses conjuntos contêm dados pessoais sensíveis; se existe representatividade demográfica adequada nas amostras de treinamento; se houve consentimento ou base legal para o uso de determinado dado; e se é possível excluir ou corrigir um dado específico em caso de solicitação de titular, sem comprometer a integridade do modelo já treinado.

Essa incapacidade de resposta não é meramente teórica. Estudos de caso amplamente discutidos na literatura sobre ética em IA — como o sistema COMPAS de avaliação de risco de reincidência criminal utilizado em cortes norte-americanas, e sistemas comerciais de reconhecimento facial com desempenho significativamente inferior para peles mais escuras — têm em comum o fato de que os desenvolvedores não possuíam, no momento do desenvolvimento, visibilidade suficiente sobre a composição demográfica e a representatividade dos dados de treinamento utilizados. Essa lacuna de visibilidade é, em essência, uma falha de Data Discovery: a ausência de um inventário confiável que permitisse identificar, previamente, o desequilíbrio amostral que viria a se manifestar como discriminação algorítmica em produção.

3.2 Descoberta de Dados e Conformidade Regulatória

A relação entre Data Discovery e conformidade regulatória pode ser observada em três frentes normativas centrais. Em primeiro lugar, a LGPD brasileira e o GDPR europeu exigem que organizações mantenham registros das operações de tratamento de dados pessoais (Records of Processing Activities, no vocabulário do GDPR), o que pressupõe conhecimento atualizado sobre quais dados pessoais existem, onde estão armazenados e com que finalidade são tratados — conhecimento esse que, em ambientes corporativos com múltiplos sistemas legados e plataformas em nuvem, só é viável em escala por meio de ferramentas automatizadas de descoberta e classificação de dados.

Em segundo lugar, o AI Act europeu estabelece, para sistemas de alto risco, a obrigação de que os conjuntos de dados de treinamento, validação e teste sejam relevantes, representativos, livres de erros na medida do possível e completos, considerando as características específicas do contexto de uso pretendido. O cumprimento dessa obrigação depende diretamente de mecanismos de descoberta e perfilamento de dados (data profiling) capazes de mensurar, de forma quantitativa e documentada, a completude, a representatividade e a qualidade dos dados de treinamento antes e durante o ciclo de vida do modelo.

Em terceiro lugar, o AI RMF do NIST e a norma ISO/IEC 42001:2023 convergem ao exigir que organizações mantenham documentação técnica detalhada sobre a origem, o

processamento e as limitações dos dados utilizados em sistemas de IA, frequentemente materializada em artefatos como data sheets for datasets — modelo de documentação estruturada de conjuntos de dados proposto na literatura acadêmica e progressivamente incorporado a práticas corporativas — e model cards, que documentam o desempenho e as limitações de modelos de aprendizado de máquina em diferentes subgrupos populacionais. A elaboração desses artefatos é, na prática, impossível sem processos maduros de Data Discovery, uma vez que exige conhecimento preciso sobre a proveniência, a composição e as características estatísticas dos dados subjacentes.

3.3 Tecnologias e Métodos Contemporâneos de Data Discovery Aplicados à IA

As soluções tecnológicas contemporâneas de Data Discovery voltadas à governança de IA articulam, tipicamente, quatro capacidades técnicas centrais. A primeira é a catalogação automatizada de dados (data cataloging), que consiste na varredura sistemática de fontes de dados estruturadas e não estruturadas para construção de um repositório centralizado de metadados, incluindo esquemas, tipos de dados, volumetria, proprietários responsáveis e classificações de sensibilidade. Plataformas contemporâneas de catalogação incorporam algoritmos de aprendizado de máquina para inferir automaticamente relações semânticas entre campos de dados dispersos em diferentes sistemas, reduzindo a dependência de documentação manual historicamente incompleta.

A segunda capacidade é a classificação de dados sensíveis (data classification), que aplica técnicas de reconhecimento de padrões, expressões regulares, dicionários semânticos e, mais recentemente, modelos de linguagem para identificar categorias de dados pessoais e sensíveis conforme taxonomias regulatórias específicas — como dados de saúde, origem racial ou étnica, orientação sexual e dados biométricos, categorias que recebem tratamento jurídico diferenciado tanto na LGPD quanto no GDPR e no AI Act.

A terceira capacidade é a linhagem de dados (data lineage), que consiste no rastreamento automatizado do fluxo de dados desde sua origem até seus pontos de consumo, incluindo todas as transformações intermediárias. Ferramentas contemporâneas de linhagem operam por meio de instrumentação de pipelines de dados (parsing de código SQL, scripts de engenharia de dados e orquestradores de fluxo de trabalho), permitindo reconstituir de forma automática e visual o caminho percorrido por um dado específico, capacidade que se revela essencial para investigações de incidentes, auditorias regulatórias e análises de causa-raiz de vieses algorítmicos.

A quarta capacidade, de consolidação mais recente e específica ao contexto de IA, é o rastreamento de metadados de aprendizado de máquina (ML metadata tracking), operacionalizado por meio de plataformas de MLOps que registram, de forma versionada, quais conjuntos de dados foram utilizados em cada experimento de treinamento, quais transformações foram aplicadas, quais hiperparâmetros foram utilizados e quais métricas de desempenho — incluindo métricas de equidade algorítmica desagregadas por subgrupo populacional — foram obtidas. Essa capacidade estende o escopo tradicional do Data Discovery, originalmente circunscrito a dados armazenados em repositórios, para o próprio ciclo de vida de desenvolvimento de modelos de IA, numa convergência técnica que a literatura recente denomina governança de dados e modelos (data-and-model governance).

3.4 Linha do Tempo Evolutiva: da Governança de Dados à Governança de Inteligência Artificial

A síntese histórica desenvolvida ao longo desta seção permite propor uma periodização evolutiva articulando governança de dados, Data Discovery e governança de IA, organizada em cinco fases sucessivas e cumulativas. A primeira fase, situada entre as décadas de 1970 e 1980, corresponde à consolidação dos sistemas de gerenciamento de banco de dados relacionais e à emergência de controles rudimentares de qualidade de dados, sem preocupação sistemática com descoberta ou catalogação.

A segunda fase, entre 1990 e 2005, corresponde à formalização da governança de dados como disciplina de gestão, com a consolidação de data warehouses, arcabouços como o DAMA-DMBOK e as primeiras ferramentas de descoberta baseadas em padrões sintáticos, ainda restritas a dados estruturados e a finalidades majoritariamente analíticas de business intelligence.

A terceira fase, entre 2005 e 2016, é marcada pela explosão do big data, pela adoção de arquiteturas de data lake e pela incorporação de processamento de linguagem natural e aprendizado de máquina às ferramentas de descoberta, ampliando sua capacidade de operar sobre dados não estruturados em larga escala.

A quarta fase, entre 2016 e 2023, corresponde à consolidação regulatória da proteção de dados pessoais, com a aprovação do GDPR e da LGPD, que transformaram o Data Discovery de conveniência analítica em obrigação de conformidade legal, ao mesmo tempo em que se iniciava, com os Princípios de IA da OCDE em 2019, a institucionalização da governança de IA como campo autônomo.

A quinta e atual fase, iniciada em 2023 e ainda em consolidação, é caracterizada pela convergência explícita entre governança de dados e governança de IA, materializada na ISO/IEC 42001:2023, no AI RMF do NIST e no AI Act europeu, todos exigindo capacidades avançadas de descoberta, catalogação, linhagem e rastreamento de metadados como pré-condição técnica para conformidade regulatória e gestão de riscos algorítmicos. Essa periodização evidencia que o Data Discovery não surgiu como resposta à governança de IA, mas sim como disciplina previamente consolidada que se reconfigurou e se tornou estrategicamente central diante das nescessidades específicas impostas pelos sistemas algorítmicos contemporâneos.

4 DISCUSSÃO: DESAFIOS, LIMITAÇÕES E PERSPECTIVAS FUTURAS

Não obstante os avanços tecnológicos e regulatórios descritos, a integração entre Data Discovery e governança de IA enfrenta desafios estruturais relevantes. O primeiro desafio refere-se à escala e à heterogeneidade dos ambientes de dados contemporâneos, caracterizados pela coexistência de sistemas legados, plataformas em nuvem de múltiplos provedores, aplicações de software como serviço (SaaS) e fontes de dados de terceiros, o que torna a construção de um inventário verdadeiramente completo e atualizado uma meta assintótica, raramente alcançada de forma plena mesmo pelas organizações mais maduras.

O segundo desafio diz respeito à limitação técnica das próprias ferramentas de descoberta automatizada diante de dados não estruturados de natureza multimodal — como combinações de texto, imagem, áudio e vídeo utilizadas no treinamento de modelos fundacionais contemporâneos —, para as quais os métodos de classificação semântica ainda apresentam taxas de erro superiores às observadas em dados estruturados tradicionais, o que introduz incerteza residual mesmo em programas de governança tecnicamente avançados.

O terceiro desafio é de natureza organizacional e cultural, e talvez o mais persistente: a governança de dados e a governança de IA frequentemente residem em estruturas organizacionais distintas — a primeira em áreas de tecnologia da informação e proteção de dados, a segunda em comitês de ética, áreas jurídicas ou unidades de ciência de dados —, o que gera lacunas de comunicação, duplicação de esforços e, em casos extremos, inventários de dados e inventários de modelos de IA que não conversam entre si, comprometendo a rastreabilidade de ponta a ponta que os frameworks regulatórios contemporâneos exigem.

O quarto desafio refere-se à tensão entre a granularidade exigida pela auditabilidade regulatória e os custos computacionais e financeiros de manter linhagem de dados e metadados

de aprendizado de máquina em escala de produção, especialmente em organizações que operam centenas ou milhares de modelos simultaneamente, cenário cada vez mais comum diante da proliferação de aplicações baseadas em modelos de linguagem de grande porte (LLMs) e de arquiteturas de IA generativa.

Como perspectivas futuras, observa-se tendência de convergência técnica entre plataformas de governança de dados e plataformas de MLOps, com o surgimento de soluções integradas que tratam o dado e o modelo como artefatos de uma mesma cadeia de proveniência unificada, abordagem que a literatura recente denomina governança unificada de dados e IA. Adicionalmente, observa-se crescente interesse em técnicas de descoberta e classificação baseadas em modelos de linguagem de grande porte, que prometem maior precisão semântica na identificação de dados sensíveis em contextos não estruturados, ainda que introduzam novos desafios relacionados à explicabilidade e à confiabilidade das próprias ferramentas de descoberta. Por fim, é razoável antecipar que futuros marcos regulatórios, no Brasil e em outras jurisdições, tenderão a explicitar, de forma cada vez mais detalhada, requisitos técnicos mínimos de descoberta, catalogação e linhagem de dados como condição de conformidade para sistemas de IA de alto risco, consolidando definitivamente o Data Discovery como infraestrutura crítica — e não mais acessória — da governança algorítmica.

5 CONSIDERAÇÕES FINAIS

Este artigo buscou demonstrar, por meio de uma análise histórico-evolutiva, que o Data Discovery constitui pilar estratégico indissociável dos programas contemporâneos de governança de Inteligência Artificial. A reconstrução da trajetória que vai da consolidação dos sistemas de gerenciamento de banco de dados relacionais, nas décadas de 1970 e 1980, até a convergência regulatória materializada no AI Act europeu, no AI Risk Management Framework do NIST e na norma ISO/IEC 42001:2023, revela um padrão histórico consistente: a capacidade de localizar, classificar e rastrear a proveniência de dados sempre precedeu, tecnicamente, a capacidade de governar os sistemas construídos a partir desses dados.

A principal contribuição deste estudo reside na sistematização de uma periodização evolutiva em cinco fases, articulando conceitos até então tratados de forma relativamente fragmentada na literatura — governança de dados, Data Discovery e governança de IA —, evidenciando sua interdependência técnica e regulatória. Argumenta-se, com base nessa periodização, que organizações que não investem em capacidades maduras de descoberta, catalogação e linhagem de dados enfrentam limitações estruturais, e não meramente

operacionais, para operacionalizar princípios éticos e cumprir exigências regulatórias de governança de IA, na medida em que a conformidade com requisitos de representatividade, qualidade, explicabilidade e auditabilidade de sistemas algorítmicos pressupõe conhecimento prévio e confiável sobre os dados subjacentes.

Como limitações deste estudo, destaca-se seu caráter eminentemente teórico e documental, sem validação empírica direta por meio de estudos de caso organizacionais aprofundados ou pesquisa de campo. Sugere-se, como agenda de pesquisas futuras, a condução de estudos empíricos que mensurem, em organizações de diferentes setores e portes, a correlação entre o grau de maturidade de Data Discovery e indicadores efetivos de conformidade e de mitigação de riscos em programas de governança de IA, bem como a investigação aprofundada de modelos de governança unificada de dados e modelos de aprendizado de máquina, campo ainda incipiente e de elevado potencial de desenvolvimento acadêmico e prático nos próximos anos.

REFERÊNCIAS

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília: Diário Oficial da União, 2018.

BRASIL. Projeto de Lei nº 2338, de 2023. Dispõe sobre o uso da Inteligência Artificial. Brasília: Senado Federal, 2023.

COMISSÃO EUROPEIA. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Bruxelas: Official Journal of the European Union, 2024.

DAMA INTERNATIONAL. DAMA-DMBOK: Data Management Body of Knowledge. 2. ed. Denville: Technics Publications, 2017.

FLORIDI, L.; COWLS, J. A unified framework of five principles for AI in society. Harvard Data Science Review, Cambridge, v. 1, n. 1, 2019.

GEBRU, T. et al. Datasheets for datasets. Communications of the ACM, New York, v. 64, n. 12, p. 86-92, 2021.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 27001:2022 — Information security, cybersecurity and privacy protection — Information security management systems — Requirements. Genebra: ISO, 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. Genebra: ISO, 2023.

JOBIN, A.; IENCA, M.; VAYENA, E. The global landscape of AI ethics guidelines. Nature Machine Intelligence, Londres, v. 1, n. 9, p. 389-399, 2019.

KITCHIN, R. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Londres: Sage Publications, 2014.

LANEY, D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Stamford: META Group Research Note, 2001.

MITCHELL, M. et al. Model cards for model reporting. In: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2019, Atlanta. Proceedings [...]. New York: ACM, 2019. p. 220-229.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. AI Risk Management Framework (AI RMF 1.0). Gaithersburg: NIST, 2023.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. Paris: OECD, 2019.

PARLAMENTO EUROPEU; CONSELHO DA UNIÃO EUROPEIA. Regulamento (UE) 2016/679, de 27 de abril de 2016 (Regulamento Geral sobre a Proteção de Dados — GDPR). Bruxelas: Jornal Oficial da União Europeia, 2016.

RUSSELL, S.; NORVIG, P. Inteligência Artificial: Uma Abordagem Moderna. 4. ed. Rio de Janeiro: GEN LTC, 2022.

ZUBOFF, S. A Era do Capitalismo de Vigilância. Rio de Janeiro: Intrínseca, 2021.