v7.n11.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

FRAMEWORK GAIA-FIS: GOVERNANÇA DE IA PARA MITIGAÇÃO DE FRAUDES POR DEEPFAKES DE ÁUDIO E IDENTIDADES SINTÉTICAS

GAIA-FIS FRAMEWORK: AI GOVERNANCE FOR MITIGATING AUDIO DEEPFAKE AND SYNTHETIC IDENTITY FRAUD

MARCO GAIA-FIS: GOBERNANZA DE LA IA PARA LA MITIGACIÓN DEL FRAUDE MEDIANTE DEEPFAKES DE AUDIO E IDENTIDADES SINTÉTICAS

Letícia Hojnacki Kologeski[1], Roberto Franciscatto[2]

e7119222
https://doi.org/10.47820/recima21.v7i11.9222

PUBLICADO: 11/2026

RESUMO

A proliferação da IA generativa criou um novo vetor de risco corporativo, *deepfakes* de áudio para fraudes financeiras. Paralelamente, o Brasil discute a regulação do setor por meio do PL nº 2.338/2023, que propõe deveres de governança. Este artigo apresenta o Framework GAIA-FIS, modelo prático de governança desenhado para organizações brasileiras. Adotou-se abordagem qualitativa, baseada em análise documental e mapeamento cruzado entre requisitos regulatórios e controles da ISO/IEC 42001:2024. Baseado no PL, na LGPD e na ISO/IEC 42001:2024, o *framework* propõe 20 controles operacionais estruturados em cinco dimensões e guiados pelo ciclo PDCA, com o propósito de reduzir a lacuna entre a conformidade legal e a defesa técnica contra identidades sintéticas. Por se tratar de uma proposta ainda não validada empiricamente, recomenda-se sua aplicação experimental em estudos futuros.

PALAVRAS-CHAVE: *Deepfakes* de Áudio. Governança de Inteligência Artificial. Fraude Financeira Corporativa.

ABSTRACT

The proliferation of generative AI has created a new corporate risk vector, audio deepfakes used for financial fraud. Concurrently, Brazil is debating the regulation of the sector through Bill No. 2.338/2023, which proposes governance obligations. This article presents the GAIA-FIS Framework, a practical governance model designed for Brazilian organizations. A qualitative approach was adopted, based on documentary analysis and cross-mapping between regulatory requirements and ISO/IEC 42001:2024 controls. Drawing on the Bill, the LGPD, and ISO/IEC 42001:2024, the framework proposes 20 operational controls structured across five dimensions and guided by the PDCA cycle, aiming to narrow the gap between legal compliance and technical defense against synthetic identities. As the proposal has not yet been empirically validated, its experimental application is recommended in future studies.

KEYWORDS: *Audio Deepfakes. Artificial Intelligence Governance. Corporate Financial Fraud.*

RESUMEN

La proliferación de la IA generativa ha creado un nuevo vector de riesgo corporativo, los deepfakes de audio utilizados para fraudes financieros. Paralelamente, Brasil discute la regulación del sector mediante el Proyecto de Ley n.º 2.338/2023, que propone obligaciones de gobernanza.

[1] Especialista em Gestão de Tecnologia da Informação pela Universidade Federal de Santa Maria - RS (UFSM).
[2] Doutor em Informática na Educação pela Universidade Federal do Rio Grande do Sul (PPGIE/UFRGS). Professor na Universidade Federal de Santa Maria - RS (UFSM).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Este artículo presenta el marco GAIA-FIS, un modelo práctico de gobernanza diseñado para organizaciones brasileñas. Se adoptó un enfoque cualitativo, basado en el análisis documental y en el mapeo cruzado entre los requisitos regulatorios y los controles de la norma ISO/IEC 42001:2024. Basado en dicho proyecto de ley, en la LGPD y en la norma ISO/IEC 42001:2024, el marco propone 20 controles operativos estructurados en cinco dimensiones y guiados por el ciclo PDCA, con el propósito de reducir la brecha entre el cumplimiento normativo y la defensa técnica frente a identidades sintéticas. Dado que la propuesta aún no ha sido validada empíricamente, se recomienda su aplicación experimental en estudios futuros.

***PALABRAS CLAVE**: Deepfakes de audio. Gobernanza de la inteligencia artificial. Fraude financiero corporativo.*

1. INTRODUÇÃO

A Inteligência Artificial (IA) consolidou-se como uma das tecnologias mais transformadoras do século XXI, redefinindo processos, otimizando operações e ampliando as fronteiras da automação corporativa. Em seu escopo mais amplo, a IA refere-se a sistemas computacionais capazes de realizar tarefas que normalmente exigiriam inteligência humana, como o aprendizado, a percepção e a tomada de decisão (RUSSELL; NORVIG, 2022). Mais recentemente, a vertente conhecida como Inteligência Artificial Generativa impulsionou essa capacidade a um novo patamar, permitindo a criação autônoma de conteúdos sintéticos inéditos e hiper-realistas, abrangendo desde a automação de textos até a geração de mídias audiovisuais de alta fidelidade (ALKHATIB, 2025).

No entanto, a popularização de ferramentas de inteligência artificial capazes de gerar áudio e vídeo realistas introduziu um risco operacional grave e novo no ambiente corporativo, a fraude por *deepfakes*. O termo *deepfake* (junção das expressões *deep learning* e *fake*) designa mídias sintéticas manipuladas e geradas por redes neurais profundas para falsificar a identidade e as características de uma pessoa real (ZHANG, 2026). No contexto organizacional, essa ameaça se materializa primordialmente por meio dos *deepfakes* de áudio, com o uso de clonagem de voz para enganar colaboradores e, por exemplo, autorizar transferências bancárias fraudulentas, modelos que já demonstram alta capacidade de evasão contra sistemas de detecção atuais (GAJEWSKA *et al.*, 2026).

Paralelamente a esse cenário de ameaças, o Brasil discute a regulação do setor por meio do Projeto de Lei nº 2.338/2023, que visa estabelecer o marco legal da Inteligência Artificial no país. Caso aprovada nos termos atuais, essa regulação poderá exigir que as empresas implementem programas de governança e gerenciem ativamente os riscos associados aos sistemas de IA que desenvolvem ou utilizam (NAKAMURA; ALMEIDA, 2026). Entre a urgência do risco material das clonagens vocais e a perspectiva da regulação estatal, surge uma lacuna

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

operacional: como as organizações corporativas podem se preparar, na prática, para traduzir o texto da lei em controles defensivos concretos?

Nesse contexto de incertezas, a adoção de um *framework* apresenta-se como uma abordagem adequada para estruturar a defesa corporativa. Esta pesquisa se justifica por buscar reduzir a lacuna supracitada, com o objetivo geral de elaborar um modelo teórico-prático de governança e *compliance* para sistemas de inteligência artificial generativa, destinado a organizações brasileiras. Para alcançar este objetivo, o presente trabalho apoia-se na convergência entre a Lei Geral de Proteção de Dados Pessoais (LGPD), o PL nº 2.338/2023 e a norma ISO/IEC 42001, que estabelece os requisitos para um sistema de gestão de IA (ABNT, 2024).

Assim, a pergunta que orienta esta pesquisa é: de que forma os requisitos da LGPD, do PL nº 2.338/2023 e da ISO/IEC 42001 podem ser traduzidos em um conjunto integrado de controles de governança voltados à mitigação de fraudes por *deepfakes* de áudio e identidades sintéticas em organizações brasileiras? Para respondê-la, foram definidos os seguintes objetivos específicos: (i) caracterizar as fraudes de identidades sintéticas e por *deepfakes* de áudio no contexto corporativo; (ii) identificar os requisitos aplicáveis na LGPD, no PL nº 2.338/2023 e na ISO/IEC 42001, distinguindo sua natureza jurídica e técnica; (iii) mapear as correspondências entre esses requisitos por meio de mapeamento cruzado (*cross-walk*); (iv) propor um catálogo de controles operacionais organizado em dimensões e articulado ao ciclo PDCA; e (v) sugerir uma estratégia de priorização e indicadores preliminares para a avaliação futura do *framework*.

2. REFERENCIAL TEÓRICO

2.1. Fraudes de Identidades Sintéticas e Deepfakes de Áudio

A evolução das tecnologias digitais tem transformado profundamente as dinâmicas de segurança cibernética, introduzindo modalidades de fraude que desafiam os mecanismos tradicionais de verificação de identidade. Nesse cenário, as fraudes de identidades sintéticas emergem como uma ameaça particularmente sofisticada, caracterizada pela fabricação integral ou parcial de personas digitais com o auxílio de inteligência artificial generativa. A literatura acadêmica define fraude de identidade sintética como a criação de uma identidade fictícia por meio da combinação de informações pessoais reais, como números de documentos válidos, com dados fabricados, resultando em uma persona que não corresponde a nenhum indivíduo real existente (IRVIN-ERICKSON, 2024). O relatório da Sumsub (2025) aponta um crescimento expressivo dessa modalidade criminosa, em que os fraudadores criam identidades sintéticas

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

completas frequentemente geradas por IA e associadas a *selfies deepfake* de alta qualidade (SUMSUB, 2025).

As fraudes de identidades sintéticas assumem múltiplas modalidades. Para montar essas identidades, utilizam-se métodos variados, incluindo:

• Falsificação inteiramente fictícia: modalidade na qual o fraudador inventa nomes, endereços e demais dados cadastrais do zero (SUMSUB, 2025).

• Uso de CPFs de crianças ou jovens: estratégia na qual o fraudador se apropria do CPF de um menor de idade para criar cadastros que acumulam histórico de crédito positivo (IRVIN-ERICKSON, 2024).

• Associação cruzada: vincula um nome fictício a um CPF legítimo de terceiros ou inativo na Receita Federal (IRVIN-ERICKSON, 2024).

• *Bust-out fraud*: o fraudador constrói gradualmente um histórico de crédito legítimo e desaparece sem quitar as obrigações (SUMSUB, 2025).

• Fraude de funcionários fictícios (*employee fraud*): geração de identidades sintéticas de trabalhadores falsos para explorar programas de benefícios corporativos (SUMSUB, 2025).

• Perfis sociais sintéticos (*synthetic social profiles*): fabricação de identidades falsas direcionadas ao ambiente de mídias e redes sociais para golpes de romance e extorsões (SUMSUB, 2025).

Após a delimitação dessas tipologias, cumpre destacar a emergência de uma categoria qualitativamente distinta: as fraudes de identidades sintéticas com apoio de *deepfakes* (BELLO *et al.*, 2026). Dentro desse espectro de ameaças, os *deepfakes* de áudio despontam como uma das manifestações mais insidiosas. A literatura acadêmica define *deepfake* de áudio como a síntese ou manipulação de voz mediada por inteligência artificial capaz de gerar enunciados coerentes e prosodicamente naturais na voz de um locutor-alvo, frequentemente sem o seu consentimento ou conhecimento (CHESNEY; CITRON, 2019).

Cabe, contudo, distinguir os dois fenômenos. Em sentido estrito, a identidade sintética corresponde a uma persona fabricada, que não remete a nenhum indivíduo real (IRVIN-ERICKSON, 2024), ao passo que o *deepfake* de áudio, em regra, falsifica a identidade de uma pessoa real e existente, como um executivo ou um fornecedor, por meio da reprodução artificial de sua voz (CHESNEY; CITRON, 2019). Para fins deste trabalho, adota-se um conceito operacional ampliado de fraude de identidade sintética, que abrange tanto (i) a fabricação de identidades fictícias quanto (ii) a representação sintética de identidades reais por meio de IA generativa, modalidade na qual se insere a clonagem vocal. Embora reconheça ambas as modalidades, o *Framework* GAIA-FIS concentra seus controles técnicos na segunda,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

especificamente nas fraudes corporativas mediadas por *deepfakes* de áudio, recorte que justifica a ênfase adotada nas seções seguintes.

A criação de áudio sintético para fraudes utiliza a inteligência artificial generativa e redes neurais profundas, que processam vastos conjuntos de dados através de modelos de aprendizado de máquina chamados *deep learning* (STRYKER; SCAPICCHIO, 2024, apud ALVES; SALANDIN; NEVES, 2025, p. 28). A clonagem vocal emprega *autoencoders* que mapeiam as características espectrais e os atributos de voz do falante para gerar uma representação de alta fidelidade, podendo resultar em áudios com elevado grau de similaridade em relação à voz original. Os índices de similaridade relatados na literatura, contudo, variam conforme a ferramenta de clonagem, a métrica empregada e as condições experimentais de cada estudo, não devendo ser interpretados como medida geral do realismo desses sistemas (PICCOLO; BORGES; FLORIAN, 2024; ALVES; SALANDIN; NEVES, 2025). A análise espectral, através de espectrogramas e modelos de detecção baseados em anomalias de fase e artefatos de síntese, é crucial para identificar a manipulação que escapa ao ouvido humano (PICCOLO; BORGES; FLORIAN, 2024).

O avanço dos Modelos de Linguagem de Grande Escala (LLMs) viabilizou a automação de ataques de engenharia social, como o *spear phishing*, eliminando erros gramaticais e aumentando a credibilidade das fraudes (POQUIVIQUI *et al.*, 2025). A materialização desses riscos já é documentada: em 2024, um gerente financeiro autorizou uma transferência de US$ 25 milhões após uma videoconferência em que todos os participantes eram *deepfakes* (LEAL; SILVA, 2025; MONTEIRO, 2025). Esse cenário culmina no conceito de vitimização sintética, em que o sofrimento penalmente relevante é causado por eventos simulados ou fabricados por IA (RAMOS; GROSS JUNIOR, 2026).

2.2. Arcabouço Normativo e Sistema de Gestão de IA

O mapeamento dos requisitos regulatórios fundamenta-se na convergência entre os preceitos de governança corporativa e a segurança da informação. Antes de apresentá-lo, contudo, cumpre distinguir a natureza dos três instrumentos considerados, que não possuem a mesma força normativa: (i) a LGPD (Lei nº 13.709/2018) constitui obrigação legal vigente; (ii) a ABNT NBR ISO/IEC 42001:2024 é norma técnica de adoção voluntária, que estabelece requisitos para sistemas de gestão de IA; e (iii) o PL nº 2.338/2023 é proposta legislativa em tramitação, cujos dispositivos ainda não produzem efeitos jurídicos. Neste trabalho, os requisitos do PL são tratados como obrigações prospectivas, e as referências a seus artigos correspondem ao texto aprovado pelo Senado Federal em 10 de dezembro de 2024.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Quanto ao seu status, em consulta realizada em outubro de 2026, o PL nº 2.338/2023 encontrava-se na Câmara dos Deputados, casa revisora, aguardando parecer do relator na Comissão Especial constituída para sua análise (BRASIL, 2026). Por se tratar de informação temporalmente sensível, eventuais alterações no texto durante a tramitação poderão modificar a numeração e o conteúdo dos dispositivos aqui referenciados.

No texto aprovado pelo Senado, o PL nº 2.338/2023 propõe uma regulação baseada em risco, prevendo que os agentes de IA implementem medidas estruturadas de governança interna, segurança desde a concepção e avaliação de impacto algorítmico, especialmente para sistemas classificados como de alto risco (BRASIL, 2023). Cabe ressaltar que os sistemas de autenticação biométrica cujo único objetivo seja confirmar a identidade de uma pessoa específica, como a verificação vocal em centrais de atendimento, foram excluídos do rol de alto risco previsto no artigo 14 do texto aprovado, de modo que o enquadramento de soluções de biometria vocal dependerá de sua finalidade e de seu contexto de uso. A LGPD, por sua vez, classifica dados biométricos como dados pessoais sensíveis (art. 5º, II), submetendo seu tratamento a um regime de proteção mais rigoroso e prevendo que a autoridade nacional poderá determinar ao controlador a elaboração do Relatório de Impacto à Proteção de Dados Pessoais (RIPD/DPIA) (art. 38) (BRASIL, 2018).

A norma ISO/IEC 42001 especifica os requisitos para estabelecer um sistema de gestão de IA no contexto de uma organização, adotando a estrutura harmonizada (HS) comum às normas de sistemas de gestão (ABNT, 2024). O ciclo PDCA (*Plan-Do-Check-Act*) orienta a interação contínua entre planejamento, operação, verificação e ação corretiva, enquanto o Anexo A da norma compila 38 controles de referência distribuídos em eixos temáticos. Embora seja de adoção voluntária, a implementação da ISO/IEC 42001 pode contribuir para a produção de evidências documentais aptas a demonstrar o cumprimento do princípio da responsabilização e prestação de contas previsto na LGPD (art. 6º, X), bem como para antecipar os deveres de governança propostos no PL nº 2.338/2023 (ABNT, 2024; BRASIL, 2018). Com base nesses instrumentos, propõem-se as correspondências operacionais sintetizadas no Quadro 1, que não configuram equivalência jurídica entre os dispositivos, mas pontos de convergência utilizados como referência na construção do *framework*:

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Quadro 1. Correspondências propostas entre a legislação brasileira (vigente e projetada)

Instrumento normativo	Dispositivo / Requisito	Controle ISO 42001	Eixo / Mecanismo de Operacionalização
PL 2.338/2023 (proposta em tramitação)	art. 3º, VII (Dever de diligência e auditabilidade).	A.6.2.8 A.6.2.7	Registro de *logs* de eventos Documentação técnica do sistema.
	art. 25 (Avaliação de impacto algorítmico).	A.5.2 A.5.4	Processo de avaliação de impacto do sistema Avaliação de impacto sobre indivíduos e grupos.
	art. 18 (Governança e segurança no ciclo de vida).	A.6.1.3 A.6.2.4	Projeto e desenvolvimento responsáveis Verificação e validação do sistema de IA.
Instrumento normativo	**Dispositivo / Requisito**	**Controle ISO 42001**	**Eixo / Mecanismo de Operacionalização**
LGPD (Lei 13.709/2018 - vigente)	art. 38 (Relatório de Impacto - DPIA).	A.5.2 A.5.3	Avaliação de impacto do sistema de IA Documentação dos resultados da avaliação.
	art. 48 (Notificação de incidentes de segurança).	A.8.4 A.8.5	Planos de comunicação de incidentes Prestação de informações às partes interessadas.
	art. 50 (Políticas de boas práticas e governança).	A.2.2 A.2.3	Implementação da política de IA Alinhamento com outras políticas organizacionais.
Transversal (Tratamento de dados / Biometria vocal)	LGPD, art. 5º, II; PL 2.338/2023, art. 17.	Eixo A.7	Salvaguardas de aquisição, proveniência e qualidade de dados para IA.

Fonte: autores.

A implantação desse sistema de gestão de IA possibilita que o uso de ferramentas algorítmicas ocorra de forma padronizada e alinhada ao cenário normativo (ABNT, 2024), o que sugere a conveniência de uma abordagem integrada, e não fragmentada, da conformidade.

3. METODOLOGIA

A literatura apresenta soluções diversas. Nakamura e Almeida (2026) focam no setor financeiro regulado pelo BACEN; Zhang (2026) propõe um *framework* para políticas de telecomunicações; Alkhatib (2025) estrutura soluções criptográficas (*watermarking* e *blockchain*);

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

e Gajewska *et al.* (2026) avaliam empiricamente detectores de *deepfakes* (NAKAMURA; ALMEIDA, 2026; ZHANG, 2026; ALKHATIB, 2025; GAJEWSKA *et al.*, 2026). Contudo, no conjunto de fontes analisado, não foram identificadas propostas operacionais voltadas diretamente para a defesa corporativa privada contra a clonagem sintética de forma integrada à legislação nacional.

Para suprir essa demanda, a pesquisa caracteriza-se como aplicada, quanto à natureza; qualitativa, quanto à abordagem; e documental e bibliográfica, quanto aos procedimentos técnicos. Seu percurso foi orientado pela lógica da *Design Science Research (DSR)*, voltada à concepção de artefatos que solucionem problemas práticos de forma fundamentada (PEFFERS *et al.*, 2007). Das seis etapas propostas por esses autores (identificação do problema, definição dos objetivos da solução, projeto e desenvolvimento, demonstração, avaliação e comunicação), este estudo contempla a identificação do problema, a definição dos objetivos, o projeto e desenvolvimento do artefato e sua comunicação, permanecendo a demonstração e a avaliação empírica como etapas futuras, conforme discutido nas limitações.

O *corpus* de análise compreendeu dois conjuntos de fontes. O primeiro, de natureza normativa, foi composto pela LGPD, pelo texto do PL nº 2.338/2023 aprovado pelo Senado Federal e pela ABNT NBR ISO/IEC 42001:2024, selecionados por constituírem, respectivamente, a principal obrigação legal vigente aplicável ao tratamento de dados biométricos, a proposta de marco regulatório da IA em discussão no país e a norma técnica de referência para sistemas de gestão de IA. O segundo, de natureza bibliográfica, reuniu literatura acadêmica e relatórios técnicos sobre fraudes de identidade sintética, *deepfakes* de áudio, detecção de voz sintética e governança de IA, selecionados de forma intencional segundo os critérios de pertinência temática ao objeto, atualidade, com predominância de publicações de 2024 a 2026, e aplicabilidade ao contexto corporativo. Por não se tratar de revisão sistemática, essa seleção não pretende exaurir a produção sobre o tema, o que constitui uma limitação do estudo.

O desenvolvimento do “Catálogo de Controles” empregou o método de mapeamento cruzado (cross-walk), executado em quatro etapas: (i) extração, de cada instrumento, dos dispositivos relacionados a governança, avaliação de riscos, segurança da informação, proteção de dados biométricos e resposta a incidentes; (ii) identificação de correspondências entre os dispositivos legais e os requisitos e controles da ISO/IEC 42001, adotando-se como critério a convergência de finalidade, isto é, a capacidade de um controle da norma operacionalizar o objetivo de proteção expresso no dispositivo legal ou proposto; (iii) associação dessas correspondências às ameaças identificadas na literatura, de modo a reter apenas os requisitos

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

pertinentes à mitigação de fraudes por *deepfakes* de áudio; e (iv) consolidação dos requisitos convergentes em controles operacionais verificáveis e rastreáveis, agrupados em cinco dimensões e articulados ao ciclo PDCA, do que resultaram os 20 controles apresentados no Quadro 2. Ressalte-se que as correspondências estabelecidas resultam da interpretação dos autores e não foram, nesta etapa, submetidas à validação externa por especialistas.

3.1. O Framework GAIA-FIS

O *Framework* GAIA-FIS estrutura-se em cinco dimensões integradas concebidas para traduzir a legislação em rotinas operacionais. A arquitetura obedece a uma progressão lógica que cobre todo o ciclo de vida do risco cibernético corporativo.

• Dimensão 1 (D1) - Governança e Responsabilidade: propõe que a governança de IA seja atribuição da alta administração. Prevê a elaboração da Política de Uso de IA Generativa, a manutenção de um inventário de sistemas de IA e a designação de um responsável formal (*AI Officer* ou comitê) para alinhar a Tecnologia da Informação (TI) aos objetivos de *compliance* da organização (ABNT, 2024; BRASIL, 2023). Ressalta-se que essa designação constitui proposta organizacional do *framework*. Embora possa articular-se com o encarregado pelo tratamento de dados pessoais previsto no art. 41 da LGPD, as funções não se confundem, uma vez que o encarregado atua no âmbito da proteção de dados pessoais, ao passo que o responsável proposto abrange, de forma mais ampla, a governança dos sistemas de IA.

• Dimensão 2 (D2) - Avaliação de Riscos de IA: prevê a execução documentada e contínua do mapeamento de riscos de IA, consolidada na Matriz de Risco. Inclui a classificação de risco dos sistemas, segundo os critérios propostos no PL nº 2.338/2023, e recomenda, como medida prudencial de governança, a elaboração do Relatório de Impacto à Proteção de Dados (RIPD/DPIA) sempre que o tratamento envolver biometria vocal. Cabe distinguir que, na LGPD, a elaboração desse relatório pode ser determinada pela autoridade nacional (art. 38), não constituindo exigência legal automática para todo tratamento de dados biométricos; sua adoção sistemática pelo *framework* decorre do elevado risco associado a esse tipo de dado pessoal sensível (BRASIL, 2018).

• Dimensão 3 (D3) - Controles Técnicos Anti-Deepfake: operacionaliza barreiras defensivas sistêmicas segundo o princípio de defesa em profundidade, no qual nenhuma camada é tratada como suficiente de forma isolada. A primeira camada consiste na inserção de APIs de *liveness detection* nos *gateways* de borda (VoIP e URAs); contudo, o desempenho desses detectores pode variar entre ambientes, degradar-se diante de ataques adaptativos e de técnicas avançadas de *spoofing*, gerar falsos positivos e criar dependência de fornecedores, razão pela qual seus resultados devem ser tratados como sinal de risco, e não como decisão definitiva

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

(GAJEWSKA *et al.*, 2026). Por isso, a detecção é complementada pela verificação fora de banda (*Out-of-Band*) acoplada a mecanismos de Autenticação Multifator (MFA), pela intervenção humana (*human-in-the-loop*) nas zonas de dúvida e pela revalidação periódica do desempenho dos próprios detectores. Completa a dimensão o registro forense das interações em repositórios imutáveis (WORM), cuja implementação deve observar os princípios de finalidade, adequação e necessidade previstos na LGPD (art. 6º, I a III): gravação restrita a canais e transações críticos, base legal definida, prazo de retenção proporcional e configurado para expiração, controle de acesso, criptografia e descarte seguro ao término do período (BRASIL, 2018). Como as gravações de voz constituem dados pessoais e, quando empregadas para identificação, podem configurar dados biométricos sensíveis, o próprio mecanismo defensivo deve ser objeto de avaliação de risco, em consonância com a privacidade desde a concepção (*privacy by design*; LGPD, art. 46, § 2º).

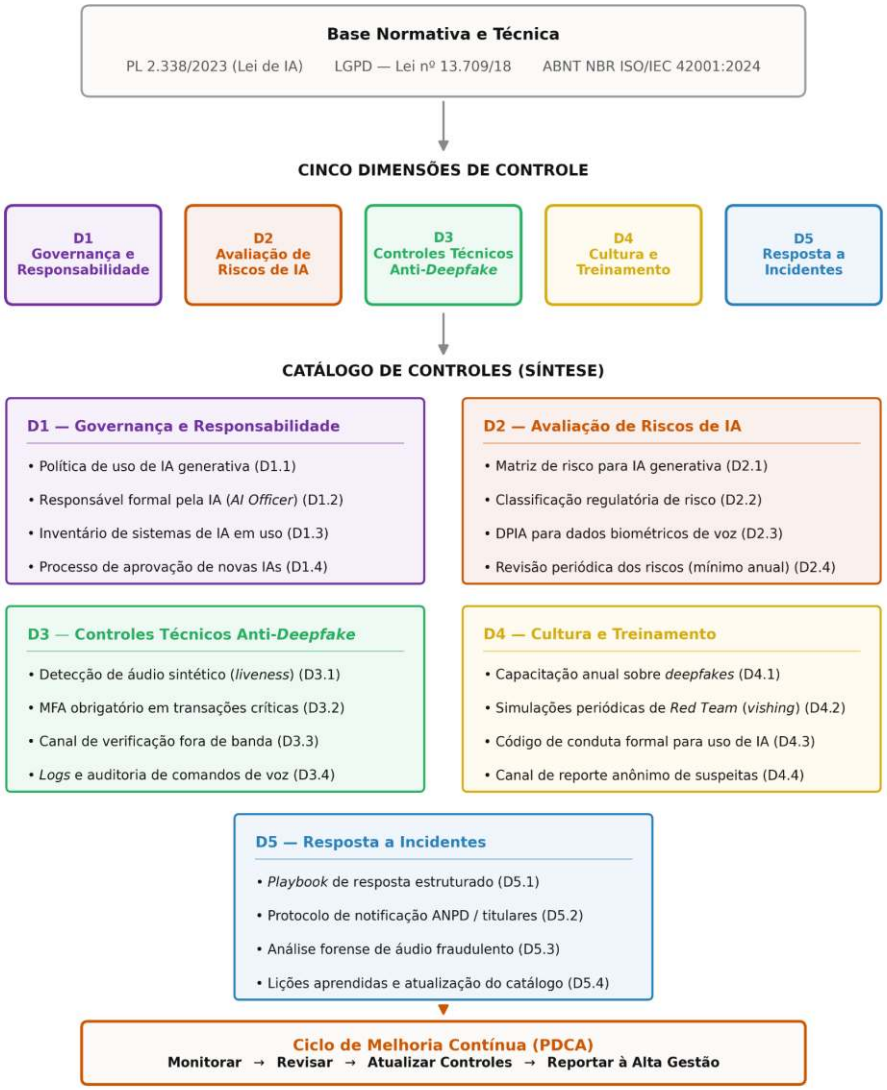
• Dimensão 4 (D4) - Cultura e Treinamento: enfrenta a vulnerabilidade do fator humano mediante capacitações periódicas. Institui o uso de simulações ativas de ataques de *vishing* (via *Red Team*) para avaliar a resiliência dos colaboradores, amparadas por códigos de conduta formais e canais de reporte (ABNT, 2024; POQUIVIQUI *et al.*, 2025).

• Dimensão 5 (D5) - Resposta a Incidentes: integra a resposta a *deepfakes* ao plano global de contingência corporativa. Inclui *playbooks* para o bloqueio via ferramentas SOAR (*Security Orchestration, Automation and Response*), protocolos de notificação legal à ANPD, condução de análise forense e incorporação das lições aprendidas (BRASIL, 2018).

A Figura 1 ilustra a arquitetura geral do modelo, demonstrando como as normativas se desdobram nas dimensões operacionais e no ciclo de melhoria contínua.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 1. Arquitetura geral do Framework GAIA-FIS



Fonte: autores.

A distribuição desses requisitos é sintetizada no Catálogo de Controles do *Framework* GAIA-FIS, detalhado no Quadro 2 abaixo, que especifica a ação esperada e a ancoragem normativa para cada item, facilitando a consulta e a auditoria durante a implementação:

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Quadro 2. Catálogo de Controles do Framework GAIA-FIS

Código	Controle	Descrição / Ação esperada	Ancoragem normativa
D1.1	Política de uso de IA generativa	Documento formal aprovado pela alta gestão com regras, limites e responsabilidades para uso de IA.	ISO 42001 §5.2; LGPD art. 50; PL 2.338 arts. 18 e 40.
D1.2	Designação de *AI Officer* ou comitê de IA	Responsável formal pela governança dos sistemas de IA, com alçada para suspender sistemas de alto risco.	ISO 42001 §5.3 e Anexo A.3.2; LGPD art. 41 (articulação com o encarregado); PL 2.338 art. 18 e 4º, XXXIV.
D1.3	Inventário de sistemas de IA em uso	Registro atualizado de todos os sistemas de IA generativa, com classificação de risco e finalidade.	ISO 42001 Anexos A.4.2 e B.4.2; LGPD art. 37.
D1.4	Processo de aprovação para novos sistemas de IA	Fluxo de *due diligence* que avalia riscos e conformidade antes da implantação de qualquer sistema.	ISO 42001 §8.1 e Anexo A.6.2.5; PL 2.338 arts. 12 e 26.
D2.1	Matriz de risco para IA generativa	Mapeamento dos riscos de fraude sintética com probabilidade, impacto e controles mitigadores.	ISO 42001 §6.1.2; PL 2.338 art. 25, § 1º.
D2.2	Classificação de risco dos sistemas	Baseada nos critérios do PL 2.338/2023: impacto em direitos fundamentais, escala e uso em decisões críticas.	ISO 42001 §6.1; PL 2.338 arts. 14 e 15.
D2.3	DPIA para tratamento de dados de voz	Relatório de impacto para sistemas que coletam, armazenam ou processam voz como dado biométrico (recomendação prudencial do *framework*; exigível quando determinado pela ANPD).	LGPD art. 38; PL 2.338 arts. 25 e 27.
D2.4	Revisão periódica dos riscos (mínimo anual)	Atualização formal da matriz de risco, incorporando novos vetores de ataque e mudanças regulatórias.	ISO 42001 §8.2 e §9.3; PL 2.338 art. 26.
D3.1	Detecção de áudio sintético (*liveness detection*)	Ferramenta automatizada para identificar padrões de voz clonada em canais de atendimento, atuando como camada de triagem combinada a D3.2 e D3.3, com revalidação periódica de desempenho e taxa de falsos positivos.	ISO 42001 Anexo A.6.2.4; PL 2.338 art. 18; LGPD art. 46.
D3.2	MFA obrigatório em transações e decisões críticas	Autenticação multifator independente do canal de voz para autorização de transferências e contratos.	ISO 42001 §8.3; LGPD art. 46.
D3.3	Verificação fora de banda para ordens por voz	Confirmação por canal alternativo (app autenticador, *token* físico ou notificação *push*), com vinculação funcional da aprovação aos dados específicos da transação (valor, destinatário e identificador único), de modo a possibilitar a unicidade e a integridade contextual da operação.	ISO 42001 §8.3; PL 2.338, arts. 17 e 18.
D3.4	*Logs* imutáveis de interações por voz	Registro auditável e inviolável das interações de voz em canais e transações críticos, limitado ao necessário à finalidade de segurança e prevenção a fraudes, com prazo de retenção definido, controle de acesso, criptografia e descarte seguro,	ISO 42001 Anexo A.6.2.8 e §7.5; LGPD arts. 6º, I a III, 37 e 46, § 2º; PL 2.338 art. 18, II, "b".

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

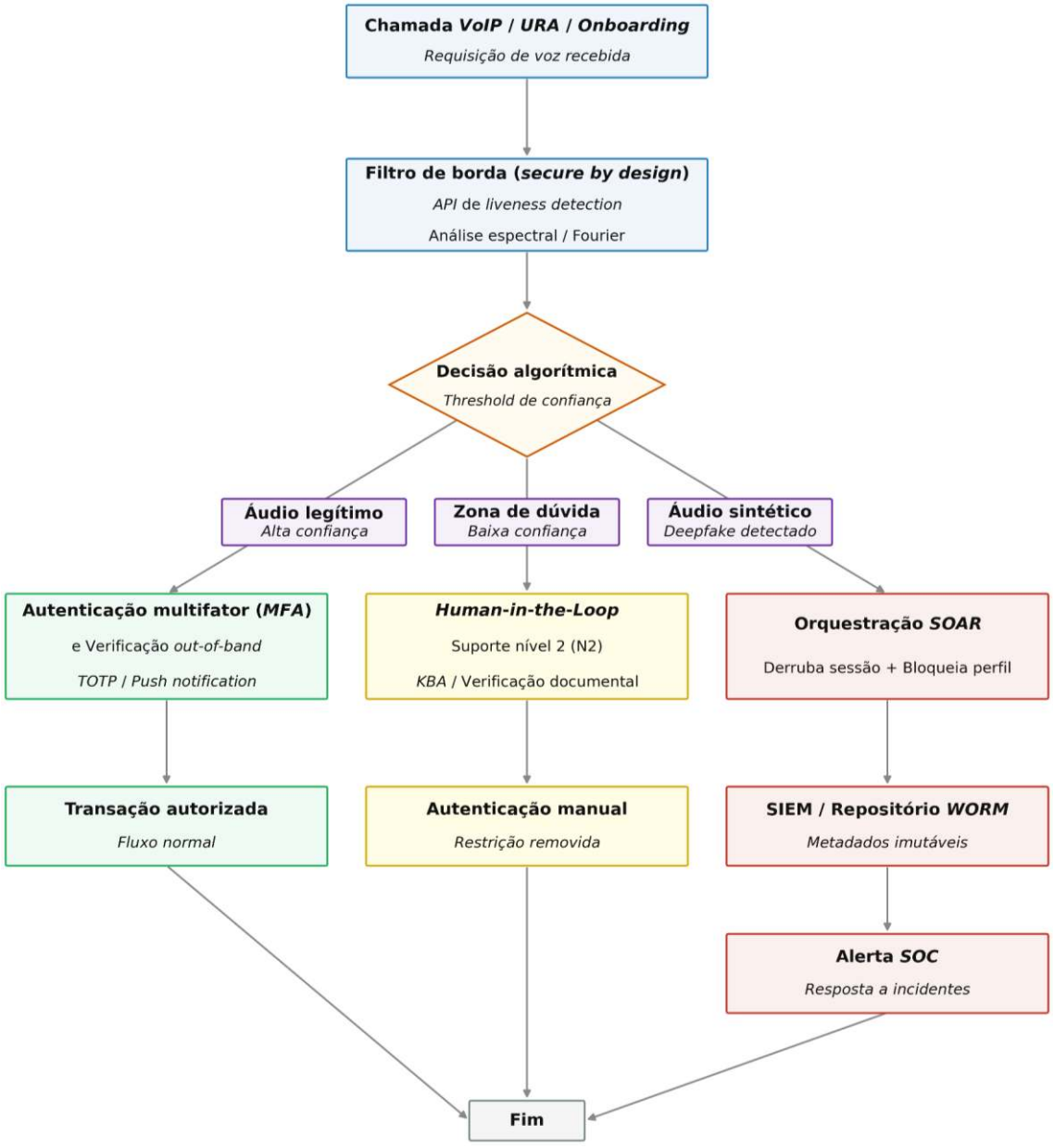
Código	Controle	Descrição / Ação esperada	Ancoragem normativa
		conforme políticas internas e exigências legais aplicáveis.	
D4.1	Capacitação anual em *deepfake*	Treinamento com exemplos reais de fraude por voz clonada, voltado a todos os colaboradores.	ISO 42001 §7.2 e §7.3; PL 2.338 art. 40.
D4.2	Simulações de ataque por voz sintética (*red team*)	Exercícios controlados com voz clonada para testar a resiliência dos colaboradores e dos processos.	ISO 42001 §7.2, §7.3 e §9.1; PL 2.338 art. 40.
D4.3	Código de conduta para uso de IA generativa	Regras claras sobre o que é permitido, vedado e como reportar uso suspeito de IA pelos colaboradores.	ISO 42001 §5.2 e Anexo A.9.2; PL 2.338 art. 40.
D4.4	Canal anônimo de reporte de suspeitas de fraude por IA	Mecanismo seguro para reportar suspeitas de *deepfake* ou uso indevido de IA sem receio de represália.	ISO 42001 Anexo A.3.3; LGPD art. 50; PL 2.338 art. 40 (programa de governança: incentivo à denúncia de irregularidades).
D5.1	Linha do tempo *(playbook)* de resposta a incidentes por *deepfake*	Procedimento com passos de contenção, investigação e comunicação para fraudes por áudio ou vídeo sintético.	ISO 42001 §10.2 e Anexo A.8.4; PL 2.338 arts. 40 (programa de governança) e 42 (comunicação de incidente grave).
D5.2	Protocolo de notificação à ANPD	Fluxo para comunicar incidentes com dados biométricos de voz à ANPD no prazo definido pela autoridade competente.	LGPD art. 48; PL 2.338 art. 42.
D5.3	Análise forense de áudio fraudulento	Capacidade interna ou contratada de perícia técnica em arquivos de áudio para identificar manipulação por IA.	ISO 42001 §10.2; LGPD art. 46.
D5.4	Lições aprendidas e atualização do *framework*	Processo formal pós-incidente- para incorporar aprendizados ao catálogo de controles e à política de IA.	ISO 42001 §10.1 e §10.2; PL 2.338 art. 40, § 2º, VI.

Fonte: autores.

A aplicação sistemática desses controles tende a favorecer que a governança de IA na organização transcenda políticas abstratas, materializando-se em salvaguardas tecnológicas auditáveis e rastreáveis. A Figura 2 detalha o fluxograma da arquitetura técnica anti-deepfake, evidenciando a tomada de decisão algorítmica, o bloqueio em tempo real e a verificação *Out-of-Band.*

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 2. Fluxograma de arquitetura técnica anti-deepfake



Fonte: autores.

4. RESULTADOS E DISCUSSÃO

Cabe esclarecer, inicialmente, que os resultados apresentados nesta seção correspondem ao artefato produzido pela pesquisa, isto é, ao modelo proposto, aos controles

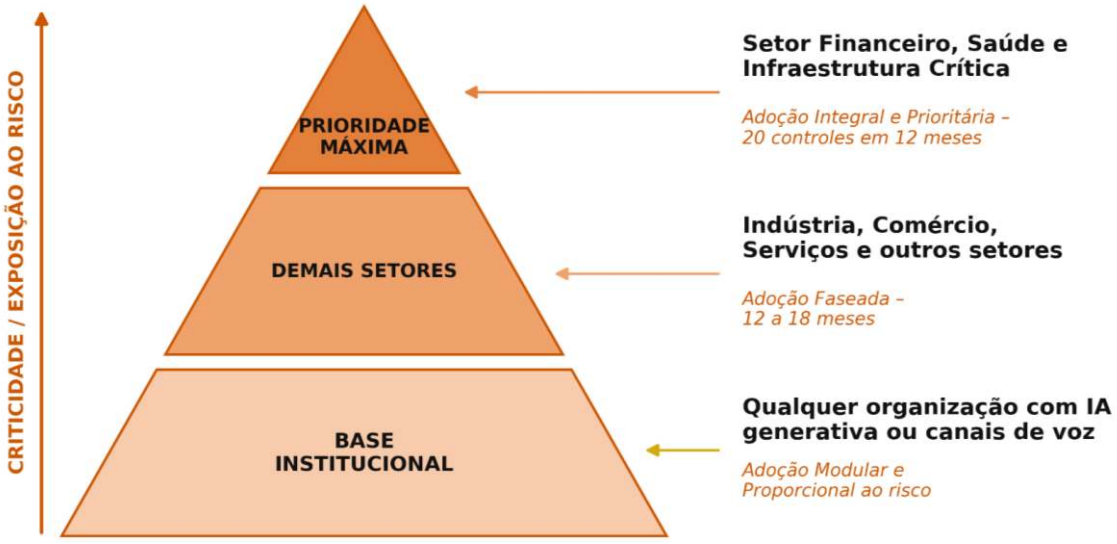
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

sugeridos e às metas preliminares, e não a resultados decorrentes de sua aplicação em organizações, etapa reservada a estudos futuros.

A arquitetura do GAIA-FIS foi concebida não como uma lista de verificação estática, mas como um sistema guiado pelo ciclo PDCA, orientado à adaptação contínua. Sugere-se que a implementação seja gradual e estruturada de forma correspondente ao perfil de risco e ao contexto de cada organização, em observância ao princípio da proporcionalidade. Para empresas de saúde, infraestrutura crítica e do setor financeiro, sugere-se, como referência preliminar, a implementação integral e prioritária dos controles do catálogo em horizonte de até 12 meses. Para organizações de menor exposição, uma abordagem faseada, ao longo de 12 a 18 meses, tende a ser mais adequada, priorizando as fundações de governança (D1) e a avaliação de riscos (D2) para, progressivamente, avançar aos controles técnicos e operacionais. Esses prazos constituem estimativas dos autores, formuladas a partir da complexidade relativa dos controles e do grau de exposição setorial, e deverão ser calibrados em aplicações empíricas, considerando a maturidade, os recursos e a infraestrutura de cada organização.

A Figura 3 apresenta a pirâmide de priorização, que organiza os três níveis de adoção do *Framework* GAIA-FIS, desde a base institucional até o topo de prioridade máxima.

Figura 3. Pirâmide de priorização da aplicação do Framework GAIA-FIS



Fonte: autores.

Esse roteiro pretende oferecer aos gestores um ponto de partida, de modo que a organização possa evoluir sua maturidade em governança de IA de forma controlada e

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

mensurável. Contudo, a aferição da efetividade do *Framework* GAIA-FIS e a comprovação da melhoria contínua (exigida pela ISO 42001) dependem da definição de métricas objetivas que traduzam o desempenho dos controles em dados quantificáveis. O Quadro 3 propõe um conjunto de indicadores (KPIs) distribuídos pelas cinco dimensões do *framework*, acompanhados de suas respectivas fórmulas, frequências e metas ilustrativas:

Quadro 3. Indicadores propostos e metas ilustrativas para o Framework GAIA-FIS

Dimensão	Indicador (KPI)	Fórmula / Métrica	Frequência	Meta Ilustrativa[1]
D1 (Governança)	Taxa de atualização do inventário de IA	(Sistemas de IA revisados no período / Total de sistemas no inventário) × 100	Trimestral	100% até o 3º mês; após, ≥ 95% de aderência.
D2 (Riscos)	Cobertura de avaliação de risco	(Sistemas com Matriz de Risco e DPIA concluídos / Total de sistemas ativos) × 100	Semestral	100% para sistemas de alto risco; ≥ 80% para médio/baixo risco.
D3 (Técnico)	Taxa de Detecção de *Deepfakes* (*True Positive Rate*)	(Ataques sintéticos bloqueados / Total de ataques simulados em testes) × 100	Mensal	≥ 95% em testes de *Red Team*.
D3 (Técnico)	Taxa de Falsos Positivos (*False Positive Rate*)	(Usuários legítimos rejeitados / Total de usuários autenticados) × 100	Mensal	< 1% (para evitar negação de serviço - DoS).
D4 (Cultura)	Taxa de resiliência em simulações de *Red Team*	(Colaboradores que resistiram ao golpe simulado / Total de colaboradores testados) × 100	Semestral	Evoluir de 60% (1º teste) para ≥ 90% (após 24 meses).
D4 (Cultura)	Tempo médio de reporte de suspeitas	Tempo decorrido (em minutos) entre a interação suspeita e o acionamento do canal D4.4.	Contínuo	< 5 minutos.
D5 (Incidentes)	Tempo Médio de Resposta (MTTR) a incidentes	Tempo entre a detecção pelo D3.1 e a contenção total pelo D5.1.	Por incidente	< 15 minutos (contenção automatizada via SOAR).
D5 (Incidentes)	Conformidade na notificação à ANPD	(Notificações enviadas dentro do prazo legal / Total de incidentes notificáveis) × 100	Por incidente	100%.

Fonte: autores

[1] Nota: as metas constituem hipóteses de desempenho propostas pelos autores, não validadas empiricamente, e devem ser calibradas por cada organização a partir de linha de base própria.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

A estrutura dos indicadores ancora-se na exigência da ISO/IEC 42001 de que a organização determine o que precisa ser monitorado e medido, bem como os métodos, a periodicidade e a análise dos resultados (§9.1), e suas fórmulas adotam métricas usuais na avaliação de sistemas de detecção, como as taxas de verdadeiros e de falsos positivos (ABNT, 2024).

Os valores-alvo, contudo, não derivam de norma técnica, de *benchmarking* setorial ou de validação empírica, constituem metas ilustrativas, isto é, hipóteses de desempenho propostas pelos autores para orientar a operacionalização do *framework*. Essa ressalva é particularmente relevante para os indicadores técnicos, uma vez que o desempenho de detectores de *deepfakes* de áudio pode cair de forma expressiva quando avaliado fora dos conjuntos de dados de referência, em condições reais de fraude (GAJEWSKA *et al.*, 2026).

Assim, a Taxa de Detecção de *Deepfakes* ≥ 95% deve ser entendida como referência a ser testada em exercícios de *Red Team*, e não como desempenho esperado de qualquer solução; o limite de 1% para a Taxa de Falsos Positivos expressa o equilíbrio desejado entre segurança e experiência do usuário; e o Tempo Médio de Resposta a Incidentes (MTTR) inferior a 15 minutos pressupõe contenção automatizada via plataformas SOAR, podendo não ser alcançável em organizações sem esse nível de automação. Recomenda-se, portanto, que cada organização estabeleça uma linha de base nos primeiros ciclos de medição e calibre as metas progressivamente, de modo que esses dados forneçam insumos para as análises críticas realizadas pela alta direção na fase de Verificação (Check) do PDCA.

No plano internacional, o GAIA-FIS dialoga com o A*rtificial Intelligence Risk Management Framework* (AI RMF 1.0) do *National Institute of Standards and Technology* (NIST), que organiza a gestão de riscos de IA em quatro funções: governar, mapear, medir e gerenciar (NIST, 2023). Há correspondência aproximada entre essas funções e as dimensões do *framework* proposto, a função governar aproxima-se de D1; mapear, de D2; medir, dos indicadores do Quadro 3; e gerenciar, de D3, D4 e D5. O perfil do AI RMF dedicado à IA generativa, por sua vez, inclui entre os riscos próprios dessa tecnologia as ameaças à integridade e à segurança da informação, e aponta a proveniência de conteúdo como área prioritária de ação (NIST, 2024). Diferentemente desses referenciais, de adoção voluntária e escopo abrangente, o GAIA-FIS propõe um recorte temático, as fraudes por *deepfakes* de áudio, e uma ancoragem explícita no ordenamento brasileiro, o que constitui sua contribuição específica, ainda que dependente de validação.

O *framework* também comporta aprofundamentos em aspectos não detalhados nesta versão. O primeiro é a gestão de fornecedores e de riscos de terceiros, uma vez que detectores de *deepfake* e plataformas de autenticação costumam ser contratados externamente, o que

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

demanda cláusulas contratuais de desempenho, auditoria e tratamento de dados. O segundo é a integração com arquiteturas de confiança zero (*zero trust*), nas quais nenhuma interação é considerada confiável apenas por sua origem aparente (ROSE *et al.*, 2020), princípio coerente com a verificação fora de banda proposta em D3.3. O terceiro são os mecanismos de proveniência e de marca d'água (*watermarking*) de conteúdos, que atuam na origem do áudio sintético e complementam a detecção (ALKHATIB, 2025). Por fim, a adoção dos controles não elimina o risco, subsiste um risco residual, que deve ser formalmente aceito pela alta direção e reavaliado periodicamente, considerando também os custos de implementação, potencialmente elevados para organizações de menor porte, o que reforça a pertinência da abordagem faseada proposta.

5. CONSIDERAÇÕES FINAIS

A popularização de ferramentas de inteligência artificial generativa introduziu no ambiente corporativo um risco operacional qualitativamente novo: a fraude por *deepfakes* de áudio. Diante das evidências de que os métodos tradicionais de segurança podem apresentar limitações relevantes diante de vozes sintéticas hiper-realistas e da perspectiva de um marco regulatório, este artigo apresentou o *Framework* GAIA-FIS com o propósito de reduzir a lacuna operacional existente. A articulação proposta entre a ISO/IEC 42001, o PL nº 2.338/2023 e a LGPD sugere que o dever de diligência não precisa se restringir a obrigações abstratas, podendo materializar-se em procedimentos concretos, verificáveis e auditáveis.

Entre as limitações da pesquisa, cabe assinalar que o *Framework* GAIA-FIS não foi submetido a um estudo de caso ou a uma aplicação empírica em organização real, o que restringe a possibilidade de mensurar a sua eficácia prática. Do mesmo modo, as correspondências normativas resultam da interpretação dos autores e não foram validadas por especialistas, as metas do Quadro 3 têm caráter ilustrativo e a seleção das fontes bibliográficas foi intencional, e não sistemática. Além disso, o escopo dos controles técnicos concentrou-se no bloqueio de ataques em canais de voz telefônicos, não abrangendo de forma homogênea a detecção multimídia em ataques por videoconferência. Por fim, como o PL nº 2.338/2023 ainda tramita no Congresso Nacional, eventuais alterações em seu texto poderão exigir ajustes no catálogo de controles proposto. Para trabalhos futuros, recomenda-se a aplicação experimental do *framework* por meio de estudos de caso, a validação dos controles e das metas por especialistas, por exemplo, com o emprego do método Delphi, e a inclusão de mecanismos de *liveness detection* multimodal.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

A análise realizada também indica que a efetividade do modelo tende a depender de duas condições basilares, o engajamento da alta direção e a alocação de recursos adequados, sem os quais a estrutura de governança tende a resultar em conformidade meramente documental. Em suma, esta pesquisa propõe que as organizações podem estruturar mecanismos de antecipação à regulação, visto que, em um cenário de ameaças que se sofisticam na mesma velocidade em que a tecnologia avança, a governança de inteligência artificial tende a se consolidar como componente estratégico de gestão de riscos e conformidade.

REFERÊNCIAS

ALKHATIB, M. A multifaceted deepfake prevention framework integrating blockchain, post-quantum cryptography, hybrid watermarking, human oversight, and policy governance. Computers, v. 14, n. 11, p. 488, 2025. DOI: https://doi.org/10.3390/computers14110488.

ALVES, J.; SALANDIN, A. T.; NEVES, J. E. D. Clonagem de voz por IA: avaliando a eficácia em ataques de phishing por voz. Revista Tecnológica da Fatec Americana, Americana, v. 13, n. 1, 2025.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). ABNT NBR ISO/IEC 42001: tecnologia da informação: inteligência artificial: sistema de gestão. Rio de Janeiro: ABNT, 2024.

BELLO, A. *et al.* Responsible AI in financial identity verification and risk mitigation. Discover Sustainability, 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União: seção 1, Brasília, DF, p. 59, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 5 out. 2026.

BRASIL. Câmara dos Deputados. Projeto de Lei n. 2.338, de 2023: ficha de tramitação. Brasília: Câmara dos Deputados, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262. Acesso em: 5 out. 2026.

BRASIL. Senado Federal. Projeto de Lei n. 2.338, de 2023. Dispõe sobre o desenvolvimento, o fomento e o uso ético e responsável da inteligência artificial com base na centralidade da pessoa humana. Texto aprovado pelo Plenário do Senado Federal em 10 dez. 2024. Brasília: Senado Federal, 2023. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 5 out. 2026.

CHESNEY, R.; CITRON, D. K. Deep fakes: a looming challenge for privacy, democracy, and national security. California Law Review, v. 107, n. 6, p. 1753-1820, 2019. DOI: https://doi.org/10.15779/Z38RV0D15S.

GAJEWSKA, J. *et al.* Audio deepfake detectors vs. real fraud: the fall of benchmarks. In: IEEE/CVF WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV), 2026. Proceedings [...]. [S. l.]: IEEE, 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

IRVIN-ERICKSON, Y. Identity fraud victimization: a critical review of the literature of the past two decades. Crime Science, v. 13, art. 3, 2024. DOI: https://doi.org/10.1186/s40163-024-00202-0.

LEAL, G. G.; SILVA, R. C. Análise forense de phishing automatizado com IA. São Paulo: Universidade Presbiteriana Mackenzie, 2025.

MONTEIRO, G. Deepfake de voz e CFOs falsos: a nova fraude corporativa de alto nível. Migalhas, [S. l.], 15 dez. 2025. Disponível em: https://www.migalhas.com.br/depeso/446123/deepfake-de-voz-e-cfos-falsos-a-nova-fraude-corporativa-de-alto-nivel. Acesso em: 5 out. 2026.

NAKAMURA, A.; ALMEIDA, T. Governança de IA em instituições financeiras: guia para conformidade com BACEN e CMN. [S. l.]: IBGIA, 2026. (IBGIA Policy Briefs).

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg: NIST, 2023. DOI: https://doi.org/10.6028/NIST.AI.100-1.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). Artificial Intelligence Risk Management Framework: generative artificial intelligence profile. NIST AI 600-1. Gaithersburg: NIST, 2024. DOI: https://doi.org/10.6028/NIST.AI.600-1.

PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A design science research methodology for information systems research. Journal of Management Information Systems, v. 24, n. 3, p. 45-77, 2007. DOI: https://doi.org/10.2753/MIS0742-1222240302.

PICCOLO, M. S. S.; BORGES, J. H. G.; FLORIAN, F. Decifrando deepfakes de áudios: uma investigação pela análise de frequência. ARACÊ, [S. l.], 2024.

POQUIVIQUI, A. D. O. *et al.* Phishing e engenharia social: a evolução do engano digital com IA generativa. In: ESCOLA REGIONAL DE INFORMÁTICA DE MATO GROSSO (ERI-MT), 14., 2025. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025.

RAMOS, V. A. de; GROSS JUNIOR, R. A nova vitimologia digital: os tipos emergentes de vítima penal na era da inteligência artificial. Revista Jurídica Gralha Azul – TJPR, [S. l.], v. 1, n. 33, 2026. DOI: https://doi.org/10.62248/qjphzf78.

ROSE, S.; BORCHERT, O.; MITCHELL, S.; CONNELLY, S. Zero trust architecture. NIST Special Publication 800-207. Gaithersburg: NIST, 2020. DOI: https://doi.org/10.6028/NIST.SP.800-207.

RUSSELL, S. J.; NORVIG, P. Inteligência artificial: uma abordagem moderna. 4. ed. Rio de Janeiro: LTC, 2022.

SUMSUB. Identity Fraud Report 2025-2026. London: Sum and Substance Ltd, 2025. Disponível em: https://sumsub.com/fraud-report-2025/. Acesso em: 5 out. 2026.

ZHANG, B. Governing deepfake realities: the truth immunity framework for telecommunications policy. Telecommunications Policy, 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.