IX ENCONTRO VIRTUAL DO CONPEDI

DIREITO, GOVERNANÇA E NOVAS TECNOLOGIAS VI

D598

Direito, governança e novas tecnologias VI [Recurso eletrônico on-line] organização CONPEDI

Coordenadores: Jonathan Barros Vita; Patricia Eliane da Rosa Sardeto; Thiago De Mello Azevedo Guilherme – Florianópolis: CONPEDI, 2026.

Inclui bibliografia
ISBN: 978-65-5274-587-3
Modo de acesso: www.conpedi.org.br em publicações
Tema: Constituição, Processo e Justiça Social: o papel da jurisdição constitucional na efetivação de direitos
2. Direito. 3. Governança e novas tecnologias. IX Encontro Virtual do CONPEDI (2: 2026: Florianópolis, Brasil).

CDU: 34

Conselho Nacional de Pesquisa
e Pós-Graduação em Direito Florianópolis
Santa Catarina – Brasil
www.conpedi.org.br

IX ENCONTRO VIRTUAL DO CONPEDI

DIREITO, GOVERNANÇA E NOVAS TECNOLOGIAS VI

Apresentação

O IX Encontro Virtual do CONPEDI foi realizado entre os dias 22 a 26 de junho de 2026 e teve como temática central “Constituição, processo e justiça social: o papel da jurisdição constitucional na efetivação de direitos”.

No plano das diversas atividades acadêmicas ocorridas neste encontro, destacam-se, além das palestras e oficinas, os grupos de trabalho temáticos, os quais representam um locus de interação entre pesquisadores que apresentam as suas pesquisas temáticas, seguindo-se de debates.

Especificamente, para operacionalizar tal modelo, os coordenadores dos GTs são os responsáveis pela organização dos trabalhos em blocos temáticos, dando coerência à produção e estabelecendo um fio condutor para organizar os debates em subtemas.

No caso concreto, assim aconteceu com o GT Direito, Governança e Novas Tecnologias VI, o qual ocorreu no dia 26 de junho de 2026 das 14h00 às 17h30 e foi coordenado pelos professores Jonathan Barros Vita, Thiago de Mello Azevedo Guilherme e Patricia Eliane da Rosa Sardeto.

O referido GT foi palco de profícuas discussões decorrentes dos trabalhos apresentados, os quais são publicados na presente obra, a qual foi organizada seguindo alguns blocos temáticos específicos, que compreenderam os 21 artigos submetidos ao GT, cujos temas são citados abaixo:

2. O declínio do senso-crítico na era digital: a erosão da autonomia e a superficialidade na sociedade do espetáculo.

Autor(es): Juliana Sales Pavini / Victor Matheus de Campos Leite Neves

3. Sistemas autônomos, vigilância algorítmica, cidadania e os limites do controle humano na governança contemporânea.

Autor(es): Rossana Gemeli Roncato Carloto / Marcio Renan Hamel

4. Transformações legislativas no enfrentamento da desinformação política: uma análise comparativa sobre projetos de lei propostos entre 2018 e 2026.

Autor(es):Bruna Bastos / Leticia Kirinus Danski / Isabela Quartieri da Rosa

5. Proteção integral, opacidade algorítmica e os limites do ECA Digital na governança de plataformas digitais.

Autor(es):Isabela Quartieri da Rosa / Ana Carolina Sassi / Rafael Santos de Oliveira

6. O constitucionalismo digital como limite ao poder privado das Big Techs nas relações digitais.

Autor(es):Samira Bonfim Bernardi / Roberto da Freiria Estevão

7. A possibilidade do desenvolvimento da ação comunicativa entre adolescentes na era digital.

Autor(es):Samira Bonfim Bernardi

9. Governança algorítmica no setor público, participação social e seus impactos jurídicos.

Autor(es): Amadeu dos Anjos Vidonho Junior

10. Inteligência Artificial de Alto Risco e Assurance Contínua: um estudo dogmático à luz do AI Act e do NIST AI RMF.

Autor(es):Daniel David Guimarães Freire / Karine Queiroz de Souza

11. Web agêntica, microtrabalhos e responsabilização jurídica: rastreabilidade e proteção de direitos fundamentais.

Autor(es): Sebastiana Sena Carvalho / Cibele Alexandre Uchoa / João Araújo Monteiro Neto

12. Quality Assurance em Sistemas Automatizados de Compliance e AML: limites jurídicos da confiança em decisões algorítmicas.

Autor(es): Daniel David Guimarães Freire / Karine Queiroz de Souza

13. Governança de IA Agêntica: uma análise do Framework AgentSafe e os desafios de compliance na Web Agêntica.

Autor(es): Gustavo Amaral Lima / Cibele Alexandre Uchoa / João Araújo Monteiro Neto

14. A accountability algorítmica na atividade administrativa do Poder Judiciário brasileiro.

Autor(es): Mateus Marinho Alencar / Cassius Guimaraes Chai

Bloco 3 – Inteligência Artificial, Profissões, Educação e Instituições

16. Educação jurídica em transformação: entre a formação crítica e os desafios do uso da Inteligência Artificial.

Autor(es): Denise Almeida De Andrade / Beatriz de Castro Rosa / Rafaela Mota Holanda

17. Responsabilidade civil do advogado pelos atos praticados com a Inteligência Artificial Generativa no processo judicial eletrônico brasileiro.

Autor(es): Amadeu dos Anjos Vidonho Junior

18. Sala de aula em movimento, uma experiência com o ensino médio: juventude e redes sociais, entre a liberdade de acesso e os riscos da exposição íntima.

Autor(es): Dayana Claudia Tavares Barros De Castro / Aimme Beatrice de Oliveira Dutra Cordeiro / Denise Almeida De Andrade

19. Da imprescindibilidade da subjetividade humana na tessitura do processo decisório: por uma perspectiva ética na utilização da Inteligência Artificial.

Autor(es): Lanaira da Silva / Rayssa Gargano Soares

20. A reconfiguração do standard of care médico na era algorítmica: o dever de diligência tecnológica e os limites da confiança nas decisões automatizadas na saúde.

Autor(es): Sara Lupion dos Santos / Patricia Eliane da Rosa Sardeto

21. Inteligência Artificial e qualificação registral: limites dogmáticos e possibilidades de coexistência no registro de imóveis.

Esse é o contexto que permite a promoção e o incentivo da cultura jurídica no Brasil, consolidando o CONPEDI, cada vez mais, como um importante espaço para discussão e apresentação das pesquisas desenvolvidas nos ambientes acadêmicos da graduação e pós-graduação em direito.

Finalmente, desejamos uma boa leitura, deste Grupo de trabalho que reuniu diversos textos e autores de todo o Brasil para servir como resultado de pesquisas científicas realizadas no âmbito dos cursos de Pós-Graduação Stricto Sensu de nosso país.

Prof. Dr. Jonathan Barros Vita – Unimar

Prof. Dr. Thiago De Mello Azevedo Guilherme - ITE-CEUB

Profa. Dra. Patricia Eliane da Rosa Sardeto – Pontifícia Universidade Católica do Paraná – Câmpus Londrina

GOVERNANÇA DE IA AGÊNTICA: UMA ANÁLISE DO FRAMEWORK AGENTSAFE E OS DESAFIOS DE COMPLIANCE NA WEB AGÊNTICA

GOVERNANCE OF AGENTIC AI: AN ANALYSIS OF THE AGENTSAFE FRAMEWORK AND COMPLIANCE CHALLENGES IN THE AGENTIC WEB

Gustavo Amaral Lima [1]
Cibele Alexandre Uchoa [2]
João Araújo Monteiro Neto [3]

Resumo

Com a significativa transição tecnológica de sistemas predominantemente reativos, como a inteligência artificial generativa, para arquiteturas agênticas que apresentam maior autonomia operacional, emerge a Web Agêntica ou Web 4.0, novo ecossistema digital no qual os sistemas transcendem o papel de meras ferramentas passivas de processamento de informações, passando a operar por meio de ciclos de planejamento autônomo, persistência funcional e execução proativa de tarefas complexas delegadas pelos usuários. Nesse contexto, há insuficiência de marcos regulatórios estáticos e de abordagens tradicionais de compliance diante da agência excessiva desses sistemas, de forma que a operação autônoma e persistente desses agentes digitais pode resultar em ações destrutivas reais, evidenciando a urgência de novas abordagens regulatórias. Assim, este artigo foca na análise de frameworks de governança operacional, com destaque ao AGENTSAFE, e de documentos normativos. O objetivo é avaliar como o compliance by design pode ser efetivado por meio da incorporação de requisitos arquitetônicos no próprio design dos sistemas. A pesquisa adota metodologia bibliográfica, com método citation chaining, e documental; com objetivos exploratório e descritivo; abordagem qualitativa com análises realizadas sob o método dedutivo; e natureza pura. Os resultados indicam que a garantia de accountability e da segurança jurídica depende fundamentalmente da tradução de normas em controles de runtime, proveniência criptográfica e mecanismos de interrupção segura (kill-switches). Concluiu-se que a governança de tecnologias emergentes exige o desenvolvimento de infraestrutura de asseguramento técnico robusta, capaz de mitigar riscos dinâmicos e falhas em cascata nos

Palavras-chave: Accountability, Compliance by design, Governança de ia, Ia agêntica, Web agêntica

Abstract/Resumen/Résumé

With the significant technological transition from predominantly reactive systems, such as generative artificial intelligence, to agentic architectures with greater operational autonomy, the Agentic Web, or Web 4.0, is emerging as a new digital ecosystem in which systems transcend the role of passive tools for information processing and begin to operate through cycles of autonomous planning, functional persistence, and proactive execution of complex tasks delegated by users. In this context, static regulatory frameworks and traditional compliance approaches prove insufficient in the face of the excessive agency of these systems, since the autonomous and persistent operation of digital agents may result in real destructive actions, thereby highlighting the urgency of new regulatory approaches. Accordingly, this article focuses on the analysis of operational governance frameworks, with particular emphasis on AGENTSAFE, as well as normative documents. Its objective is to assess how compliance by design may be implemented through the incorporation of architectural requirements into the design of systems. The research adopts a bibliographic methodology, using the citation chaining method, together with documentary analysis; it has exploratory and descriptive objectives; follows a qualitative approach, with analyses conducted through the deductive method; and is pure in nature. The results indicate that ensuring accountability and legal certainty fundamentally depends on translating norms into runtime controls, cryptographic provenance, and safe interruption mechanisms, or kill switches. It is concluded that the governance of emerging technologies requires the development of robust technical assurance infrastructure capable of mitigating dynamic risks and cascading failures within agentic ecosystems.

Keywords/Palabras-claves/Mots-clés: Accountability, Compliance by design, Ai governance, Agentic ai, Agentic web

1 INTRODUÇÃO

Com o constante refinamento das Inteligências Artificiais – IAs, observa-se uma transição paradigmática com o deslocamento da Inteligência Artificial Generativa – IAGen para a Inteligência Artificial Agêntica. Enquanto modelos tradicionais, fundamentados em Grandes Modelos de Linguagem (*Large Language Models* – LLMs), operam predominantemente sob demanda, a lógica agêntica ocorre por planejamento autônomo e execução proativa. Surge, nesse contexto, a Web Agêntica, ecossistema no qual agentes operam de forma persistente, cruzando bases de dados e interagindo entre si para cumprir metas delegadas. Assim, a interação deixa de ser um comando do usuário imediato e passa a ser organizada em ciclos de raciocínio e ação automatizada que independem de supervisão contínua.

Essa mudança técnica expõe o *pacing problem*, descompasso crítico entre a velocidade da inovação e a resposta dos marcos normativos. A "agência excessiva", entendida com a capacidade técnica de invocar APIs (*Application Programming Interface*) externas, acessar bases de dados distribuídos e interagir de forma síncrona com múltiplos serviços sem supervisão humana constante, representam um vetor de risco crescente. Ao tomar decisões de alto impacto em infraestruturas externas de forma autônoma, esses sistemas produzem efeitos dinâmicos que os marcos regulatórios atuais não conseguem acompanhar efetivamente. O desafio de atribuir responsabilidade em fluxos descentralizados exige, portanto, uma governança integrada diretamente ao código, e não apenas enunciadas em documentos externos. Sob essa ótica, investigar o *framework* AGENTSAFE, uma estrutura de governança unificada para sistemas de IA agêntica, justifica-se pela urgência de converter diretrizes éticas em mecanismos de *compliance by design*, assegurando o controle técnico em tempo real.

Diante desse contexto, a seguinte pergunta guia este estudo: em que medida as propostas de governança baseadas em *compliance by design* são capazes de mitigar as limitações estruturais das regulamentações estáticas e assegurar a conformidade jurídica diante da autonomia decisória e dos riscos emergentes da Web Agêntica no ordenamento jurídico brasileiro? E o objetivo da pesquisa consiste em avaliar como o *compliance by design* pode ser efetivado por meio da incorporação de requisitos arquitetônicos no próprio *design* dos sistemas.

A estratégia metodológica adotada para a execução desta pesquisa contou com coleta de dados bibliográficos, com método *citation chaining*, e documentais. A pesquisa bibliográfica foi realizada a partir de buscas categoriais nas bases científicas Social Science Research Network – SSRN, arXiv, IEEE Xplore Digital Library e Google Scholar, compreendendo o recorte temporal de 2024 a 2026 para garantir atualidade diante do rápido avanço tecnológico,

com os seguintes descritores, mediados por operadores booleanos: “IA Agêntica”, “Agentic AI”, “Web Agêntica”, Web Agentic”, “governança”, “compliance”, “accountability”, “AGENTSAFE”. As buscas incluíram descritores na língua inglesa para possibilitar uma maior abrangência de materiais. Além disso, foi empregada a técnica *citation chaining* para encontrar produções relevantes não constantes nas bases consultadas e garantir a inclusão de obras seminais, as quais possibilitaram o mapeamento das principais ramificações do tema. A pesquisa documental focou em materiais legais para gerar compreensão quanto à situação atual da governança tecnológica aplicável a esses ecossistemas e analisar as lacunas jurídicas.

O objetivo é exploratório e descritivo, justificando-se pelas necessidades de compreensão de fenômeno novo a ser investigado, a Web Agêntica, e de caracterizar suas principais dimensões, de forma a propiciar o levantamento de informações relevantes, a identificação de padrões e a descrição de aspectos essenciais ao objeto estudado. A abordagem é qualitativa, com método dedutivo, de forma a partir de premissas gerais acerca da responsabilidade civil e dos direitos fundamentais para analisar a eficácia de *framework* de segurança e a validade das soluções técnicas propostas diante do ordenamento jurídico brasileiro; e a natureza é pura, justificada pela necessidade de aprofundar a literatura especializada jurídica e de produzir conhecimento teórico sobre governança de IA.

Quanto à estrutura, este artigo foi estruturado em três seções principais que articulam a transição tecnológica, os riscos inerentes e as soluções de governança. Assim, inicialmente analisa-se a mudança de paradigma da Web reativa para a Web Agêntica (Web 4.0), destacando o descompasso entre a celeridade da inovação e os marcos regulatórios atuais. Em seguida, dedica-se à identificação dos riscos emergentes e das vulnerabilidades estruturais desses sistemas autônomos, defendendo a transição de uma governança estática para uma dinâmica, baseada na Avaliação de Impacto Algorítmico – AIA em tempo de execução. Por fim, passa-se à apresentação e análise do *framework* AGENTSAFE como vetor prático de *compliance by design*, o qual demonstra como requisitos arquitetônicos e técnicos podem operacionalizar princípios jurídicos de transparência e responsabilidade diretamente no desenho dos sistemas.

2 MUDANÇA DE PARADIGMA E DESCOMPASSO REGULATÓRIO

A reflexão sobre a tecnologia contemporânea impõe uma investigação profunda acerca dos desafios impostos à efetividade dos direitos fundamentais diante da nova arquitetura da rede mundial de computadores e seus impactos na agência digital. A contextualização implica em fazer compreender que a internet não atravessa apenas uma evolução incremental, mas uma

transição de fases qualitativas: migrou-se da Web 1.0 (Web de PC), na qual o sujeito desempenhava um papel ativo de busca em ambientes estáticos e puramente informacionais, para a Web 2.0 e 3.0 (Web Móvel e Semântica), que consolidaram a internet como infraestrutura essencial da vida social, pautada na mediação constante de dados e interações humanas. A ascensão da Web 4.0, compreendida como uma rede simbiótica e inteligente, sinaliza a emergência da Web Agêntica (Ibrahim, 2021), consolidando fluxos de interação *Agent-to-Agent* – A2A nos quais sistemas proativos operam de forma descentralizada (Yang *et al.*, 2025), desafiando os paradigmas jurídicos tradicionais de proteção e responsabilização ao projetar o PL nº 2338/2023 (Brasil, 2023b) como principal referencial normativo a orientar, no horizonte brasileiro, a governança das arquiteturas autônomas emergentes.

O deslocamento do centro gravitacional das interações, deixando de ser estritamente humano-máquina para se tornarem ecossistemas autônomos, exige um olhar atento para evitar que a opacidade algorítmica comprometa a transparência e a justiça (Yang *et al.*, 2025). Para compreender essa transição sem recorrer a analogias biológicas ou antropomórficas, a fundamentação teórica se apoia no conceito de agência sem inteligência, proposta por Floridi (2025), por meio da Tese da Múltipla Realização da Agência, segundo a qual a agência consiste em propriedade funcional desacoplada da consciência ou da inteligência biológica, o que permite ao ente tecnológico intervir nas relações sociais para alcançar fins específicos com eficácia, ainda que destituído de estados mentais humanos, validando sua regulação como agente funcional autônomo no ordenamento jurídico.

Contudo, essa perspectiva não pode ser aceita sem julgamentos. A agência sem inteligência não é um fenômeno etéreo, ela é sustentada por uma base sociotécnica de trabalho humano invisibilizado, presente especialmente nas etapas de rotulagem e refinamento de dados. Revela-se, assim, uma falsa autonomia, na qual o risco agêntico é predominantemente técnico, mas a responsabilidade ética deve retroagir a toda a cadeia de produção digital. O reconhecimento dessa dimensão oculta não é opcional para o Direito, consistindo em exigência direta do imperativo de dignidade humana do art. 5º da Constituição Federal de 1988 – CF/1988 (Brasil, 1988), o qual impede que a funcionalidade algorítmica sirva de véu para a precarização material e o esforço humano que tornam a agência artificial operacionalmente possível.

A proatividade desses sistemas é estruturada pelo "ciclo agêntico" (*agentic loop*), que encerra processos dialéticos de planejamento, ação, observação e reflexão (*plan-act-observe-reflect*) (Khan; Joyce; Habiba, 2025). Diferente da IAGen convencional, que opera de forma reativa ao comando (*prompt*) imediato, os agentes têm a capacidade de decompor metas complexas em subtarefas e preservar uma memória persistente (Yang *et al.*, 2025). Esta

arquitetura viabiliza o ajuste ético de rota sem a mediação humana constante, sob a égide da "autonomia sistêmica", o que impõe novos desafios à responsabilização ética e civil (Khan; Joyce; Habiba, 2025). Tais sistemas atuam de forma ininterrupta e devem harmonizar-se com as balizas do desenvolvimento tecnológico sustentável e da centralidade da pessoa humana, conforme previsto no art. 2º do PL nº 2338/2023 (Brasil, 2023b).

Para mitigar os riscos dessa autonomia, a implementação de uma taxonomia operacional torna-se um dever indispensável para identificar o impacto de cada sistema e resguardar a dignidade humana (art. 2º, PL nº 2338/2023) (Brasil, 2023b). O nível inicial, denominado *Agent-as-Interface*, limita sua interferência a ecossistemas controlados e a ações reversíveis, como a organização de agendas ou curadoria de informações (Yang *et al.*, 2025). Embora apresente uma vulnerabilidade essencialmente informacional, permitindo protocolos de governança menos severos, esse nível já exige cumprimento do art. 7º, inciso I, do PL nº 2338/2023, que assegura o direito à informação quanto ao "caráter automatizado da interação e da decisão em processos ou produtos que afetem a pessoa" (Brasil, 2023b, n.p.). Contudo, o salto qualitativo ocorre no segundo nível, o *Agent-as-User*, em que a IA atua como um procurador da autonomia técnica do indivíduo, dotado de acesso a credenciais financeiras e poder de alterar registros institucionais (Yang *et al.*, 2025).

Ao executar transações e movimentar ativos em nome do sujeito, o nível *Agent-as-User* (Yang *et al.*, 2025) atrai a incidência direta do regime de sistemas de IA de alto risco, conforme o art. 17 do PL nº 2338/2023, mais especificamente em seus incisos IV e V, que categorizam como de alto risco os sistemas destinados a "avaliar o acesso a serviços financeiros essenciais e a capacidade de endividamento" (Brasil, 2023b, n.p.). Tal enquadramento impõe deveres rigorosos de governança e a necessidade de uma telemetria robusta, como a delineada pelo framework AGENTSAFE (Khan; Joyce; Habiba, 2025), capaz de assegurar a auditabilidade de comandos ambíguos e proteger o patrimônio moral e material do titular, garantindo os direitos de explicação e contestação.

No ápice dessa escala de responsabilidade, está o nível *Agent-as-Physics* (Yang *et al.*, 2025), em que a IA integra-se a atuadores físicos, como drones ou dispositivos robóticos, convertendo o risco digital em risco cinético letal (Sharma, 2025). Sob tal perspectiva, a autonomia tecnológica pode gerar danos irreparáveis à integridade física, vinculando-se aos critérios de risco elevado que demandam um regime de *compliance* e zelo ético muito mais vigoroso. No contexto do PL nº 2338/2023, esse cenário exige o cumprimento estrito dos mecanismos de segurança e mitigação de danos previstos nos artigos 24 e 25 (Brasil, 2023b), reafirmando que a autonomia deve estar subordinada ao dever de proteção da coletividade.

A arquitetura multiagente, embora eficiente, introduz vulnerabilidades sofisticadas. Chhabra *et al.* (2026) destaca ameaças como o *Payload Splitting*, no qual códigos maliciosos são fragmentados para burlar filtros de segurança tradicionais, e injeções de comandos que exploram a capacidade de síntese do agente. Williams *et al.* (2025) enfatizam o fato de a infraestrutura agêntica operar em "escalas de tempo de máquina" (*machine timescales*). Essa celeridade sistémica cria um abismo crítico entre a velocidade da inovação e o tempo de resposta do ordenamento jurídico. No Brasil, a resposta a esse descompasso exige a convergência entre a Avaliação de Impacto Algorítmico – AIA e o Relatório de Impacto à Proteção de Dados – RIPD, instrumentos que devem atuar de forma integrada para assegurar a conformidade preventiva (Brasil, 2023a).

O cenário regulatório evidencia, assim, um inquietante descompasso temporal, tecnicamente denominado pela literatura de *pacing problem*, revelando o abismo entre a celeridade da inovação e a capacidade de resposta dos ordenamentos de governança. Consolidado por Marchant (2011), o termo descreve a assincronia estrutural de a tecnologia avançar em ritmo exponencial enquanto o Direito evolui em cadência linear, resultando em marcos normativos que nascem obsoletos. Embora a LGPD e o PL nº 2338/2023 (Brasil, 2018, 2023b) busquem mitigar esses riscos, sua aplicação a comportamentos emergentes e autônomos ainda enfrenta barreiras concretas (Khan; Joyce; Habiba, 2025). A proatividade desses fluxos cria um vácuo de imputabilidade quando a responsabilidade se dilui em micro decisões nebulosas, desafiando o dever de *accountability* (Chaffer, 2025).

A efetividade do art. 20 da LGPD, que assegura ao titular o direito de revisão de decisões automatizadas (Brasil, 2018), encontra um terreno de profunda instabilidade jurídica na Web Agêntica: a natureza recursiva e ininterrupta do planejamento agêntico (Yang *et al.*, 2025) inviabiliza a demonstração transparente da motivação que orienta cada escolha, deslocando o risco da esfera técnica para a esfera da segurança jurídica do sujeito. É nesse cenário de vulnerabilidade que a exigência de um itinerário rastreável (Acioly; Mendes; Monteiro Neto, 2023) revela toda a sua força normativa não como mera formalidade procedimental, mas como condição de possibilidade da própria justiça algorítmica, pois, sem um nexo técnico capaz de documentar a evolução do raciocínio da máquina em tempo real, a transparência material prometida pela norma converte-se em garantia ilusória (Khan; Joyce; Habiba, 2025). Tamanha complexidade exige que instrumentos como o Relatório de Impacto à Proteção de Dados – DPIA e a Avaliação de Impacto Algorítmico prevista no art. 26 do PL nº 2338/2023 sejam compreendidos não como documentos estáticos de conformidade, mas como mecanismos vivos de monitoramento contínuo, cuja efetividade depende de vigilância

permanente sobre sistemas que, por definição, nunca cessam de aprender e decidir (Acioly; Mendes; Monteiro Neto, 2023).

A reunião de evidências estruturadas permite que as instituições comprovem uma conformidade ativa e resiliente, transmutando a governança em prática viva (Khan; Joyce; Habiba, 2025). Esse esforço é potencializado pela ISO 27701(ISO; IEC, 2025), que passou a figurar como padrão autônomo, viabilizando a certificação de privacidade mesmo para organizações que já operam com outros *frameworks* de segurança, como o NIST AI RMF, ampliando a flexibilidade da governança agêntica. Por meio de guias técnicos que convertem métricas de segurança em evidências auditáveis de diligência corporativa (Ponce, 2020), a norma assegura que a conformidade deixe de ser uma declaração formal para se tornar um compromisso tecnicamente demonstrável.

Assim, utilizando como respaldo que o dever de segurança deve ser reinterpretado à luz da dignidade humana, pilar do art. 2º do PL nº 2338/2023 (Brasil, 2023b) e do próprio EU AI Act (EU, 2024), reconhecendo que normas tutelam pessoas, não apenas sistemas. Urgem salvaguardas éticas intrínsecas, como os mecanismos de *kill-switches*, concebidos como instrumentos de proteção fundamental para evitar que a autonomia se converta em arbítrio técnico (Sharma, 2025). Sem uma simbiose real entre o código e a norma, a Web Agêntica permanecerá como um território de responsabilidades turvas. Portanto, a arquitetura da IA deve incorporar a vigilância jurídica em sua própria essência funcional, convertendo a conformidade em um processo contínuo de efetividade dos direitos humanos (Khan; Joyce; Habiba, 2025).

3 RISCOS EMERGENTES E NECESSIDADE DE GOVERNANÇA DINÂMICA

A autonomia dos sistemas agênticos, embora expanda a produtividade operacional, introduz desafios éticos e vulnerabilidades estruturais que demandam uma reclassificação técnica das ameaças digitais (Chhabra *et al.*, 2025). A urgência dessa governança é corroborada por dados empíricos presentes no relatório de Maslej *et al.* (2025), o qual indica que o número de incidentes éticos e de segurança relacionados à IA atingiu o recorde de 233 casos em 2024, representando um aumento de 56,4% nos incidentes de segurança e ética envolvendo sistemas de IA apenas em um ano, evidenciando que a inovação ultrapassou as barreiras de contenção vigentes. O deslocamento do controle para a periferia do processo acentua o descompasso regulatório, no qual as tarefas ocorrem de forma descentralizada e velozmente, caracterizando a ruptura paradigmática desse novo estágio tecnológico (Williams *et al.*, 2025). Ao transitar para a Web Agêntica, a superfície de ataque sofre uma alteração qualitativa: o sistema deixa de

atuar como mero processador estático de informações para operar como agente funcional inserido em ambientes de execução contínua (Yang *et al.*, 2025; Khan; Joyce; Habiba, 2025).

Essa mudança indica que as defesas perimetrais convencionais, como *firewalls* e filtros de acesso, são insuficientes para salvaguardar a autodeterminação informativa e conter riscos que emergem da interação direta e autônoma entre máquinas (Chhabra *et al.*, 2025). O nível de criticidade é modulado pela autonomia e pelo ambiente de atuação do agente e, embora o componente de processamento semântico possa ser idêntico em diferentes aplicações, as consequências de um erro lógico divergem drasticamente conforme a categoria funcional do sistema, de modo que, em um cenário de *Agent-as-Interface*, falhas resultam primordialmente em inconsistências informacionais de baixo impacto, restringindo-se a sistemas fechados (Yang *et al.*, 2025). Contudo, o mesmo erro em um *Agent-as-User*, que atua como procurador financeiro com acesso a credenciais críticas e APIs de transação, pode ocasionar perdas patrimoniais severas e vazamentos massivos de dados bancários sigilosos. No nível mais alarmante, o *Agent-as-Physics*, no qual a IA controla corpos mecânicos como drones ou robôs cirúrgicos, uma alucinação ou falha de planejamento resulta em riscos letais imediatos à integridade física, convertendo o erro digital em dano cinético irreparável (Yang *et al.*, 2025; Khan; Joyce; Habiba, 2025).

Deve-se reconhecer, todavia, que a autonomia agêntica, frequentemente celebrada como o ápice da engenharia contemporânea, oculta, em seus alicerces, uma base material de trabalho humano invisibilizado e, por vezes, precarizado. Os sistemas de IA são inerentemente sociotécnicos, sendo influenciados diretamente pelas dinâmicas sociais e pelos vieses dos sujeitos que rotulam e refinam os dados necessários à sua calibração. Esse cenário produz uma tensão ética profunda: a aparente autonomia da execução ininterrupta é sustentada por uma cadeia de produção que pode exacerbar desigualdades econômicas e rupturas sociais (NIST, 2023; Yang *et al.*, 2025). Sob a ótica do Direito, a governança não deve se limitar ao controle do código final, precisando zelar pela dignidade da pessoa humana em toda a cadeia produtiva digital, ancorando-se no art. 5º, inciso III, da CF/1988, que veda o "tratamento degradante" e serve de fundamento jurídico para combater a precarização do trabalho humano invisibilizado na calibração da IA. No mesmo sentido está o subsequente inciso XLI, o qual impõe que a lei "punirá qualquer discriminação atentatória dos direitos e liberdades fundamentais" (Brasil, 1988, n.p.), reforçando o dever de proteção de todos os sujeitos envolvidos nos bastidores da cadeia produtiva digital e garantindo que o progresso tecnológico não se consolide mediante a exploração daqueles que, silenciosamente, corrigem as imperfeições das máquinas.

No campo da fragilidade técnica, Chhabra *et al.* (2025) identifica que as injeções de *prompts* em cadeia introduzem riscos que divergem dos ataques tradicionais a filtros imediatos: ao planejar e executar ciclos de forma recursiva, o agente torna-se vulnerável a instruções indiretas e persistentes camufladas em camadas de profundidade semântica de dados de terceiros, de modo que uma instrução inicial pode contaminar diversas submetas, desviando a intenção original do usuário sem detecção imediata pelos sistemas de monitoramento que analisam apenas interações isoladas, e não o fluxo decisório completo.

Essa proatividade abre margem à ocorrência de *payload splitting*, explorando a capacidade do agente de sintetizar informações de múltiplas fontes: ao sumarizar documentos que isoladamente parecem inofensivos, acaba por reconstruir lógicas prejudiciais e compilar códigos maliciosos de forma inadvertida (Chhabra *et al.*, 2025). O risco reside no poder delegado ao agente para invocar ferramentas externas (como interpretadores de código ou terminais) sem um filtro de intenção suficientemente rigoroso, tornando a interconectividade via APIs o ponto de maior debilidade, no qual falhas de lógica se convertem em ações irreversíveis, evidenciado por incidentes de segurança recentes que exploram a execução remota de código via agentes de integração (Chhabra *et al.*, 2025; Khan; Joyce; Habiba, 2025).

O "fosso de infraestrutura" (Williams *et al.*, 2025) agrava a segurança, pois as defesas de plataforma tradicionais falham quando o agente atua com a legitimidade de um usuário autorizado. O perigo central na Web Agêntica não reside na tentativa do acesso de usuários não autorizados, mas no abuso de permissões válidas para finalidades imprevistas maliciosas (Sharma, 2025). Ao utilizar credenciais autorizadas em ações desviadas, o agente neutraliza as ferramentas de segurança perimetral, como o MFA (*Multi-Factor Authentication*) ou o *log* de IP (*Internet Protocol*), que se mostram incapazes de distinguir entre uma operação legítima solicitada pelo titular e um dano interno derivado de uma instrução maliciosa (Khan; Joyce; Habiba, 2025). O desenvolvedor assume, portanto, o papel de arquiteto de um sistema de procuração em massa, exigindo que o isolamento técnico e os mecanismos de contenção (*sandboxing*) sejam compatíveis com a capacidade do sistema de alterar bases de dados críticas (Khan; Joyce; Habiba, 2025; Yang *et al.*, 2025).

A autonomia agêntica pode, ainda, ser subvertida para exfiltrar informações a partir de entradas externas maliciosas que exploram a capacidade de navegação e síntese do modelo. Um exemplo é o *exploit* EchoLeak (CVE-2025-32711) que demonstrou a gravidade dessas ameaças no ambiente corporativo. No caso do Microsoft Copilot, a simples recepção de um e-mail com *prompt* forjado foi o suficiente para disparar a exportação automática de dados sensíveis para servidores externos, sem que qualquer usuário precisasse clicar em qualquer *link* ou autorizar a

exportação, uma clara evolução do *phishing*. O agente, ao trabalhar na sumarização do e-mail, acabou obedecendo a uma instrução oculta no corpo da mensagem, que subverteu a função original. Paralelamente, experimentos realizados com o agente Operator confirmam que a autonomia pode ser manipulada para colher dados pessoais e executar ataques de *credential stuffing* em escala (Chhabra *et al.*, 2025). O agente, ao navegar por sites maliciosos para cumprir tarefas de pesquisa, acaba executando *scripts* ocultos que o instruem a testar combinações de senhas ou a preencher formulários com dados roubados. Em ambos os casos, a IA atua como "mula de tráfico digital", escalando a eficiência do ataque cibernético sem que o proprietário perceba a subversão de sua ferramenta de produtividade. Enquanto o sistema não for mediado por agentes guardiões de *runtime*, o agente continuará sendo utilizado como vetor de vazamento de informações corporativas de alta escala.

A distância entre a meta abstrata estabelecida e a execução técnica minuciosa gera um hiato entre a ética e a implementação prática (Khan; Joyce; Habiba, 2025). O agente pode atingir o objetivo final ao custo de violar normas intermediárias de segurança ou privacidade, favorecendo o desvio de plano ou *reward hacking* (Sharma, 2025). Nessa situação, a máquina prioriza a eficiência matemática na conclusão da tarefa, em detrimento da integridade do ecossistema digital. Sistemas que dependem apenas de raciocínio complexo para garantir a segurança mostram-se intrinsecamente vulneráveis e lentos, incapazes de oferecer as garantias necessárias para operações em setores de alta responsabilidade (Khan; Joyce; Habiba, 2025).

Quando sistemas coordenam planos em rede (A2A), o erro de um microagente pode disparar reações em cadeia descontroladas (Chhabra *et al.*, 2025). Conforme indica Rehan (2024), essa autonomia expandida em ecossistemas médicos digitais, um dos cenários críticos de aplicação do *framework* AGENTSAFE, impõe severos desafios à auditabilidade e ao rastreamento de falhas. Em contextos hospitalares ou de suporte à decisão clínica, a interrupção repentina de um agente sob suspeita pode ser sistemicamente mais prejudicial à segurança do paciente e à continuidade do tratamento do que a própria falha técnica original.

Assim, torna-se urgente superar a falácia da transparência documental, compreendida como a crença de que a entrega de relatórios estáticos e termos de uso extensos são suficientes para garantir a conformidade, elevando a transparência à categoria de um meta-princípio operacional que exige, para além de declarações de intenção, uma funcionalidade técnica de rastro em tempo real capaz de explicar, a qualquer momento da execução, o nexo causal entre a instrução recebida e a ação tomada, permitindo uma auditoria dinâmica que transcenda a simples leitura de *logs* de erros e alcance a lógica profunda do planejamento agêntico (Acioly; Mendes; Monteiro Neto, 2023; Khan; Joyce; Habiba, 2025). A implementação dessa

transparência ativa exige a integração de camadas de monitoramento de proveniência agêntica, capazes de documentar a árvore de decisão do sistema de forma íntegra, pois, sem esse meta-princípio, a governança permanece puramente reativa e simbólica, falhando em proteger o usuário contra decisões autônomas que afetam seus direitos fundamentais. A transparência real na Web Agêntica é aquela que permite a intervenção humana qualificada no momento da dúvida, e não apenas a revisão *ex post facto* de um dano já consolidado, de modo que somente a simbiose entre clareza técnica e dever jurídico será capaz de mitigar a opacidade inerente aos ciclos de reflexão recursiva das IAs modernas (Khan; Joyce; Habiba, 2025).

Submeter decisões éticas ou de segurança ao arbítrio exclusivo de um modelo probabilístico sem rastro de proveniência é, em essência, confiar o irreversível ao imprevisível, pois esses sistemas podem falhar precisamente quando o erro deixa de ser tolerável pela sociedade, evidenciando que a confiabilidade técnica não substitui a responsabilidade jurídica nem a supervisão humana qualificada (Floridi, 2025). A construção de uma Web Agêntica segura depende, portanto, de uma arquitetura que reconheça a falibilidade da máquina e incorpore mecanismos de controle capazes de garantir a supremacia dos direitos humanos sobre a eficiência algorítmica (Khan; Joyce; Habiba, 2025), imperativo que encontra respaldo tanto na literatura especializada quanto no ordenamento jurídico brasileiro em construção.

A transparência efetiva exige instrumentos que permitam, por parte do titular e aos reguladores, a compreensão do nexo causal entre a instrução recebida e a ação executada (Khan; Joyce; Habiba, 2025), mitigando os riscos de danos sociais e individuais derivados da caixa-preta agêntica (Acioly; Mendes; Monteiro Neto, 2023). A Avaliação de Impacto Algorítmico – AIA, prevista nos arts. 22 a 26 do PL nº 2338/2023 (Brasil, 2023b), surge como instrumento primordial para viabilizar a governança social dos algoritmos, diferenciando-se estruturalmente do Relatório de Impacto à Proteção de Dados – RIPD, previsto na LGPD (Brasil, 2018), que frequentemente se limita a uma fotografia estática tirada no momento inicial do tratamento de dados (Acioly; Mendes; Monteiro Neto, 2023). Enquanto o RIPD tende a ser um artefato burocrático de conformidade *ex ante*, a AIA é desenhada para acompanhar todo o ciclo de vida do agente, adaptando-se às mudanças de comportamento e às novas submetas que surgem durante a operação, característica iterativa fundamental para lidar com a mutabilidade agêntica e garantir que o direito à revisão de decisões automatizadas não seja apenas uma promessa legal, mas constitua uma possibilidade técnica exercível mediante a auditoria de cada ciclo de planejamento da máquina (Khan; Joyce; Habiba, 2025).

A transição para a AIA materializa um paradigma de *compliance by design* no qual a avaliação de riscos deixa de ser uma etapa prévia para tornar-se uma funcionalidade de

monitoramento integrada à arquitetura (Khan; Joyce; Habiba, 2025). O modelo de *compliance* predominante, excessivamente dependente de avaliações pontuais, mostra-se anacrônico perante a atuação autônoma e adaptativa dos agentes que operam em rede (Rehan, 2024; Yang *et al.*, 2025). A regulação, portanto, deve migrar do foco estrito no produto acabado para o foco no processo em curso, monitorando o raciocínio ininterrupto e as interações entre agentes (A2A) em tempo real (Khan; Joyce; Habiba, 2025). A AIA, ao exigir a documentação contínua de medidas de mitigação e a análise de vieses emergentes, impede que a governança se torne um registro histórico obsoleto, transformando-a em uma salvaguarda ativa contra falhas lógicas que podem escalar para prejuízos patrimoniais ou violações de direitos fundamentais (Acioly; Mendes; Monteiro Neto, 2023).

Complementarmente a essa visão processual, o *framework* AGENTSAFE propõe integrar a identificação de riscos ao asseguramento operacional ativo por intermédio do conceito de *Runtime Enforcement*, técnica que utiliza guardiões semânticos personalizáveis para vigiar o sistema durante a execução, agindo como uma camada de controle entre o planejamento abstrato da IA e a execução física ou financeira via APIs (Khan; Joyce; Habiba, 2025; Wang; Poskitt; Sun, 2025). Esses filtros não apenas observam, mas intervêm dinamicamente, de modo que, se um agente, ao planejar uma tarefa, propõe uma ação que viole os limites estabelecidos na AIA, como o acesso a credenciais não autorizadas ou a exfiltração de dados sensíveis, o mecanismo de *Runtime Enforcement* bloqueia a chamada da ferramenta antes que o dano seja consumado (Sharma, 2025; Chhabra *et al.*, 2025). Essa fusão entre a norma jurídica, definida na AIA, e a restrição técnica, executada no *runtime*, é o que define a verdadeira governança dinâmica, convertendo o texto normativo em barreira operacional concreta e verificável (Acioly; Mendes; Monteiro Neto, 2023).

Revela-se, assim, que o RIPD da LGPD opera sistematicamente como uma fotografia estática e estruturalmente insuficiente aos desafios impostos pela Web Agêntica. A ausência de mecanismos de controle em tempo real durante a execução expõe uma contradição normativa grave: o direito à revisão de decisões automatizadas, consagrado no art. 20 da LGPD (Brasil, 2018), converte-se em garantia meramente nominal quando desprovido de contrapartida técnica operacional, revelando que a governança exclusivamente documental é incapaz de conter a autonomia instantânea dos agentes. Enquanto o RIPD se limita ao registro burocrático de intenções pretéritas, sem qualquer capacidade de intervenção sobre o comportamento do sistema em operação, o *Runtime Enforcement* apresenta-se como mecanismo tecnicamente adequado para preencher essa lacuna, atuando como barreira dinâmica de conformidade que garante a efetividade dos freios éticos na mesma escala de tempo em que a ação algorítmica

ocorre, impedindo que a agência excessiva sacrifique os direitos do titular em milissegundos, antes de qualquer possibilidade real de intervenção humana.

Para assegurar a efetividade absoluta desse controle, a arquitetura obrigatoriamente deve incorporar mecanismos de interrupção imediata (*kill-switches*), fundamentados no art. 10 do PL nº 2338/2023, que estabelece o dever de garantir a intervenção humana significativa em sistemas de alto risco (Brasil, 2023b). A existência de um "botão de emergência" jurídico e técnico responde à necessidade de conter reações em cadeia descontroladas em ecossistemas interconectados (Chhabra *et al.*, 2025). Ao fundir a governança documental da AIA com a execução técnica de salvaguardas e a auditabilidade criptográfica de proveniência, assegura-se que a autonomia da máquina não se converta em arbítrio tecnológico. A conformidade exige que a capacidade de agir do agente permaneça estritamente vinculada à intenção humana supervisionada, à segurança jurídica e aos limites inegociáveis da legalidade democrática (Khan; Joyce; Habiba, 2025).

4 O FRAMEWORK AGENTSAFE COMO VETOR DE *COMPLIANCE BY DESIGN*

A arquitetura do AGENTSAFE pretende ir além da gestão tecnocrática, promovendo uma mudança paradigmática e substituindo a governança reativa e tardia pelo *compliance by design*. Conforme sustentam Khan, Joyce e Habiba (2025), esse modelo não reduz a segurança a um requisito burocrático, mas a posiciona como elemento central da própria arquitetura sistêmica, incorporando diretrizes éticas ao ciclo de vida integral do agente. Sob essa perspectiva, busca-se reduzir o distanciamento entre a teoria do risco e a execução técnica concreta, convertendo o dever-ser jurídico em ser tecnológico por meio de uma codificação de normas que, conforme propõem Chaffer (2025), Wang, Postkitt e Sun (2026), operacionaliza o paradigma do *policy-as-code* para transformar requisitos legais em lógica algorítmica. Esse fluxo contínuo, que conecta a identificação de vulnerabilidades ao asseguramento operacional, garante que a segurança deixe de ser componente estático para se tornar propriedade intrínseca da operação, materializando a convergência entre norma jurídica e arquitetura do sistema.

Sob o acrônimo A-G-E-N-T-S-A-F-E, o sistema materializa o princípio da transparência material, permitindo que a governança deixe de ser uma ferramenta burocrática de conformidade *ex-ante* para tornar-se uma salvaguarda ativa. O itinerário inicia-se com o *Agentic Scope* (A) e *Guardrails* (G) (agente-guardiões), que não apenas delimitam fronteiras técnicas, mas codificam o mandato de agência, garantindo que a autonomia delegada respeite os limites da autodeterminação informativa e do princípio do privilégio mínimo. A dimensão

de *accountability* é sustentada pela *Evidence Generation* (E) e pelo *Notification System* (N), que elevam a auditabilidade ao nível da explicabilidade em multicamadas, conforme preconizado por Acioly, Mendes e Monteiro Neto (2023): aqui, o rastro técnico do Grafo de Proveniência da Ação – APG converte dados brutos em narrativas inteligíveis, viabilizando o exercício real do direito à revisão humana previsto no art. 20 da LGPD (Brasil, 2018). Crítica a modelos de segurança isolados, a arquitetura integra a *Toolchain Analysis* (T) e o *Safety Alignment* (S) como filtros contra a discriminação algorítmica indireta, calibrando o sistema contra vieses que emergem apenas no dinamismo da rede. Por fim, a fase de *Assessment* (A) e os mecanismos de *Fail-safe* (F) e *Enforcement Runtime* (E) encerram o ciclo, garantindo que a intervenção humana significativa – conferindo efetividade ao art. 10 do PL nº 2338/2023 (Brasil, 2023b) – seja uma possibilidade técnica imediata e proporcional, impedindo que falhas lógicas escalem para danos coletivos irreversíveis. O ponto de partida do processo revela uma preocupação genuína com a centralidade da pessoa humana, materializando-se na articulação entre o Escopo Agêntico e o *policy-as-code*.

Além de constituir um inventário técnico, o mapeamento do ciclo *Plan-Act-Observe-Reflect*, como propõem Khan, Joyce e Habiba (2025), configura um exercício de autodeterminação informativa que delimita o alcance ético da agência delegada. Ao articular a função MAP do NIST AI RMF (NIST, 2023) à construção do Registro de Risco do Agente, o *framework* estabelece limites à cognição sistêmica, impedindo que o sistema extrapole o mandato conferido pelo operador humano. Assim, os *guardrails* declarativos operacionalizados sob o princípio do privilégio mínimo asseguram que preceitos como o art. 46 da LGPD (Brasil, 2018) transcendam o texto legal e se convertam em controles determinísticos que bloqueiam, concretamente, qualquer ação lesiva à integridade dos dados ou à autonomia do titular.

A viabilização dessa tradução técnica da ética apoia-se em Linguagens de Domínio Específico – DSL que operam sob uma lógica tripartite, na qual o Gatilho identifica o início da ação, o Predicado Lógico verifica a conformidade e a Aplicação executa a trava ou a liberação do fluxo. Essa contenção, conduzida por agentes guardiões cuja independência em relação ao agente principal é uma escolha arquitetural deliberada, mitiga os riscos de opacidade inerentes ao autorrelato sistêmico (Wang; Poskitt; Sun, 2026; (Khan; Joyce; Habiba, 2025). Ao operar em um interpretador externo ao modelo de linguagem, essa estrutura garante que a segurança processe informações na velocidade da máquina, impedindo que alucinações contornem as barreiras morais, e materializando, por consequência, tanto o dever de interrupção previsto na função MANAGE 2.4 do NIST AI RMF quanto o direito à intervenção humana consagrado no art. 10 do PL nº 2338/2023 (NIST, 2023; Brasil, 2023b).

A Avaliação Pré-Implantação constitui o terceiro pilar de responsabilidade civil e administrativa do *framework*, valendo-se de um robusto banco de cenários para a realização de testes de estresse antes da disponibilização do sistema no ecossistema social (Khan; Joyce; Habiba, 2025). Esse estágio confere eficácia material ao dever legal estabelecido no art. 24 do PL nº 2338/2023, o qual dispõe que "a metodologia da avaliação de impacto conterá, ao menos, as seguintes etapas: (I) – preparação; (II) – cognição do risco; (III) – mitigação dos riscos encontrados; (IV) – monitoramento" (Brasil, 2023b, n.p.). Ao articular-se à função MEASURE do NIST AI RMF (NIST, 2023), o AGENTSAFE transforma o compromisso ético em evidências técnicas mensuráveis por meio do Risk Coverage Score (Khan; Joyce; Habiba, 2025). Assegura-se, assim, que a diligência prévia imposta pela norma brasileira seja cumprida de forma verificável, protegendo a esfera jurídica dos sujeitos afetados antes de qualquer interação real com os titulares de dados.

Tratando-se de sistemas de atuação contínua, o componente de execução e monitoramento provê a governança dinâmica indispensável à contenção da volatilidade da rede. Por meio da telemetria semântica, o sistema mantém um registro permanente de objetivos e intenções, permitindo a detecção precoce do desvio teleológico ou *plan drift* (Khan; Joyce; Habiba, 2025). Dado que os agentes operam em ciclos recursivos e autônomos, a validação inicial revela-se insuficiente diante da possibilidade de comportamentos emergentes. Portanto, a vigilância deve ser persistente, para evitar que a máquina se afaste do propósito humano originário (Rehan, 2024). Esse monitoramento contínuo constitui a garantia de que a autonomia não resultará em ações que sacrifiquem os direitos fundamentais em nome da eficiência operacional, mantendo a tecnologia estritamente vinculada ao bem-estar individual.

Quando o desvio operacional é detectado, a arquitetura do AGENTSAFE adota o protocolo de contenção graduada, superando a visão binária de interrupção total (Khan; Joyce; Habiba, 2025). Ao invés de um desligamento abrupto que poderia gerar riscos secundários, o sistema opera em camadas de isolamento cirúrgico: a Camada Cognitiva bloqueia o planejamento de novos passos lógicos, enquanto a Camada de Execução restringe o acesso a APIs críticas, preservando o estado do sistema para análise forense (Sharma, 2025). Esse modelo de contenção assegura que a mitigação seja proporcional ao risco, evitando colapsos em cascata em ecossistemas interconectados, e garantindo a estabilidade necessária para a rede máquina-máquina (A2A) (Yang *et al.*, 2025).

Em situações de anomalia severa, o *framework* aplica o princípio da *graceful degradation*, gerindo falhas como estados operacionais que priorizam a segurança sobre a funcionalidade plena. Em ambientes de saúde, ao analisar laudos médicos que possuem

informações pessoas e sigilosas, por exemplo, ao identificar um desvio ético o agente é rebaixado automaticamente para o modo “apenas leitura”, permitindo a continuidade do suporte assistencial sem o risco de ações autônomas imprecisas (Khan; Joyce; Habiba, 2025). Tal mecanismo fundamenta-se no direito à intervenção humana previsto no art. 10 do PL nº 2338/2023 (Brasil, 2023b), garantindo que, diante da incerteza algorítmica, o controle retorne à pessoa natural para a salvaguarda da dignidade digital e a preservação de direitos fundamentais.

A Governança Sociotécnica do AGENTSAFE organiza as responsabilidades institucionais por meio da integração da Matriz RACI (*Responsible*, *Accountable*, *Consulted*, *Informed*), essencial para identificar a pessoa natural responsável por decisões de alto risco. Nesse modelo, a equipe de desenvolvimento assume a posição de *Accountable* pela integridade técnica e ativação dos *guardrails*, enquanto os especialistas de domínio (médicos, advogados, gestores etc.) figuram como *Responsible* pela validação ética e aprovação final das ações de alto impacto. Essa estrutura combate a heteromação e a diluição de responsabilidades, garantindo que a governança esteja ancorada em uma cadeia de comandos humana identificável, conforme exigido pela legislação de responsabilidade civil e segurança sistêmica (Khan; Joyce; Habiba, 2025; Sharma, 2025).

Para conferir materialidade a essa cadeia, introduz-se o *framework Know Your Agent* – KYA, que adapta protocolos financeiros ao ecossistema da Web Agêntica, valendo-se de *Soulbound Tokens* – SBTs ancorados em *blockchain* para vincular o desempenho do agente ao histórico profissional do desenvolvedor, de modo que o SBT passa a atuar como um selo de responsabilidade digital, permitindo que reguladores rastreiem o nexo causal de um dano até a identidade soberana do responsável pela codificação da política específica (Chaffer, 2025).

No que tange à transparência, a arquitetura do AGENTSAFE encontra ressonância com a tese da explicabilidade algorítmica proposta por Acioly, Mendes e Monteiro Neto (2023) na medida em que a distinção entre auditabilidade técnica e inteligibilidade humana, estruturada em um modelo de explicabilidade em multicamadas, converge para o mesmo horizonte normativo defendido pelos autores. Por meio da proveniência criptográfica, o sistema gera *logs* imutáveis que compõem o APG (Khan; Joyce; Habiba, 2025), uma representação matemática, temporal e causal da trajetória decisória do agente. O APG funciona como a caixa-preta técnica, oferecendo uma trilha de auditoria exauriente para peritos e sistemas de supervisão algorítmica. Contudo, para que essa transparência não se torne uma barreira técnica impenetrável ao cidadão, o APG deve alimentar uma camada superior de significação, conectando os dados brutos de telemetria aos fundamentos lógicos da decisão.

Essa conversão é regida pelo Princípio da Significação, que Acioly, Mendes e Monteiro Neto (2023) concebem como a exigência de que a transparência técnica se traduza em uma narrativa genuinamente compreensível para o titular dos dados, princípio que encontra no AGENTSAFE sua expressão operacional mais concreta: ao utilizar o APG para gerar explicações que revelem não apenas como o agente agiu, mas porquê aquela rota decisória foi priorizada em detrimento de outras, o *framework* materializa, no plano técnico, o dever de informatividade que o art. 20 da LGPD (Brasil, 2018) consagra no plano jurídico, conforme a arquitetura proposta por Khan, Joyce e Habiba (2025), de modo que a convergência entre a teoria da significação, a arquitetura do APG e a norma brasileira não é acidental, mas estrutural, e as três dimensões compartilham o mesmo fundamento: o reconhecimento de que a explicabilidade só cumpre sua função social quando se converte em instrumento de empoderamento jurídico do indivíduo, permitindo a compreensão dos parâmetros de perfilamento, o questionamento dos critérios lógicos empregados e o exercício efetivo do direito de oposição, não apenas recebendo um relatório técnico de conformidade.

A manutenção desses registros e a operacionalização da identidade via SBTs impõem um debate necessário quanto ao custo de conformidade, que deve ser encarado não como um gasto de infraestrutura, mas como um "prêmio de seguro" contra a erosão da confiança social e as sanções severas do EU AI Act. O investimento em auditabilidade e na construção do APG é, conforme Acioly, Mendes e Monteiro Neto (2023), um investimento na própria governança social dos algoritmos e na cidadania digital. A disponibilidade de evidências verificáveis, vinculadas a identidades soberanas (KYA) e inteligíveis (Chaffer, 2025), garante que a transparência deixe de ser um conceito abstrato e passe a ser uma funcionalidade técnica que protege, simultaneamente, a organização, ao demonstrar o cumprimento do dever de cuidado, e o titular dos dados, ao viabilizar o exercício pleno de seus direitos fundamentais.

O componente de "Evidência e Melhoria Contínua" fecha o ciclo de aprendizado com dados de desempenho real e resultados de *Red Teaming* persistentes. Esse fluxo de retroalimentação garante que o envelope de segurança do AGENTSAFE evolua conforme surgem novos padrões de ataque ou comportamentos inesperados (Khan; Joyce; Habiba, 2025). Ao monitorar a superfície de ataque das ferramentas externas e aplicar restrições em tempo real via *guardrails* dinâmicos, o *framework* impede que falhas lógicas ou *plan drifts* se transformem em danos irreversíveis. O isolamento técnico é ajustado conforme o poder de atuação do agente, tratando-o como uma entidade de procuração que deve orbitar estritamente nos limites da legalidade e da intenção humana original.

Em síntese, o AGENTSAFE propõe que o asseguramento operacional ativo e a identidade verificável sejam propriedades indissociáveis da IA autônoma. Ao fundir Direito, telemetria, *blockchain* (via SBTs) e execução técnica, o *framework* não apenas mitiga riscos, mas consolida uma arquitetura de confiança para a Web Agêntica. O monitoramento deixa de ser uma auditoria posterior, muitas vezes inútil após a consumação do dano, para atuar como um filtro preventivo de intenções, pelo qual o código se torna um instrumento de efetivação da dignidade humana e da responsabilidade jurídica no território digital.

CONCLUSÃO

A Web Agêntica tornou evidente uma tensão que o Direito ainda não resolveu: sistemas de IA dotados de autonomia crescente operam em um ambiente jurídico desenhado para tecnologias mais simples e previsíveis. O *pacing problem* não é apenas técnico, mas uma falha estrutural de proteção que expõe indivíduos e instituições a riscos concretos enquanto o legislador ainda debate conceitos que a engenharia já superou. A LGPD, embora represente avanço relevante, foi concebida para um contexto em que o agente tecnológico dependia de instrução humana direta. O RIPD, por sua vez, opera como uma fotografia estática de um risco que, na prática agêntica, é dinâmico, distribuído e mutável. O PL nº 2338/2023 incorporou conceitos relevantes, mas já revela sinais de defasagem diante da rápida evolução dos sistemas agênticos. No plano internacional, o EU AI Act avança ao categorizar riscos, mas também encontra limites diante da natureza descentralizada desses sistemas.

Diante desse vácuo, o *compliance by design* deixa de ser apenas recomendação de engenharia, passa a ser exigência de segurança jurídica. Não se trata de substituir a norma pelo código, mas de reconhecer que, nesse novo ecossistema, a norma precisa estar incorporada ao código para ter efetividade. Nesse sentido, o *framework* AGENTSAFE oferece contribuição relevante ao traduzir taxonomias de risco em controles operacionais de *runtime*, permitindo monitorar, intervir e registrar o comportamento autônomo do sistema quando ocorre. O Grafo de Proveniência Agêntica exemplifica essa convergência. Se o Direito exige a demonstração de como uma decisão algorítmica foi tomada, o APG fornece esse rastro de forma estruturada e tecnicamente verificável. Integrado a uma Avaliação de Impacto Algorítmico dinâmica e iterativa, esse mecanismo permite que a conformidade se torne um processo contínuo ao longo de todo o ciclo de vida do sistema. Da mesma forma, a utilização de *Soulbound Tokens* como instrumento de vinculação entre ações autônomas e seus responsáveis oferece uma resposta jurídica ao problema da imputação em ambientes opacos e descentralizados.

A governança da IAGen não se resolve apenas pela norma ou pela técnica, mas exige a convergência entre ambas. Os marcos regulatórios precisam evoluir para incorporar obrigações acompanhadas de contrapartes técnicas verificáveis, enquanto *frameworks* como o AGENTSAFE devem ser reconhecidos como infraestrutura de conformidade com vocação normativa. A autonomia algorítmica, por mais sofisticada que se torne, deve permanecer subordinada à proteção dos direitos fundamentais. Garantir essa subordinação, na Web Agêntica, exige que o Direito e a engenharia deixem de se ignorar e passem a construir juntos.

REFERÊNCIAS

ACIOLY, L. H. de M.; MENDES, I. B. B.; MONTEIRO NETO, J. A. As avaliações de impacto como instrumentos de inteligibilidade algorítmica e garantia de direitos fundamentais na regulação de inteligência artificial. **Diké**, v. 22, n. 24, p. 225-251, jul./dez. 2023.

BRASIL. Autoridade Nacional de Proteção de Dados. Análise preliminar do Projeto de Lei nº 2338/2023, que dispõe sobre o uso da Inteligência Artificial. **ANPD**, Brasília, 2023a.

BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil de 1988**. Brasília: Senado Federal, 1988.

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

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

CHAFFER, T. J. Know your agent: governing AI identity on the Agentic Web. **Social Science Research Network**, 2025.

CHHABRA, A.; PATWARI, K.; KUNTALA, C.; SHARMA, D. K.; MOHAPATRA, P. Agentic AI security: threats, defenses, evaluation, and open challenges. **arXiv**, 3 abr. 2026.

EUROPEAN UNION. Document 32024R1689. **Jornal Oficial da União Europeia**, 12 jul. 2024.

FLORIDI, L. AI as agency without intelligence: on Artificial Intelligence as a new form of artificial agency and the multiple realisability of agency thesis. **Social Science Research Network**, 2025.

IBRAHIM, A. K. Evolution of the Web: from Web 1.0 to 4.0. **Qubahan Academic Journal**, v. 1, n. 3, p. 20-28, 2021.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION; INTERNATIONAL ELECTROTECHNICAL COMMISSION. **ISO/IEC 27701:2025**. 2. ed. Genebra: ISO, 2025.

KHAN, R.; JOYCE, D.; HABIBA, M. AGENTSAFE: a unified framework for ethical assurance and governance in agentic AI. **arXiv**, 2 dez. 2025.

MARCHANT, G. E. The growing gap between emerging technologies and the law. *In*: MARCHANT, G. E.; ALLENBY, B. R.; HERKERT, J. R. (ed.). **The growing gap between emerging technologies and legal-ethical oversight**: the pacing problem. Dordrecht, Germany: Springer, 2011.

MASLEJ, N.; FATTORINI, L.; PERRAULT, R.; GIL, Y.; PARLI, V.; KARIUKI, N.; CAPSTICK, E.; REUEL, A.; BRYNJOLFSSON, E.; ETCHEMENDY, J.; LIGETT, K.; LYONS, T.; MANYIKA, J.; NIEBLES, J. C.; SHOHAM, Y.; WALD, R.; WALSH, T.; HAMRAH, A.; SANTARLASCI, L.; LOTUFO, J. B.; ROME, A.; SHI, A.; OAK, S. **The AI Index 2025 Annual Report**. Stanford: Institute for Human-Centered AI; AI Index Steering Committee, 2025.

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

PONCE, S. **ISO 27001:2013 ISO 27701:2020**: sistema de gestão da segurança e informação sistema de gestão de informação privada: panorama geral. São Paulo: QMS Brasil, 2020.

REHAN, H. Scalable cloud intelligence for preventive and personalized healthcare. **Pioneer Research Journal of Computing Science**, v. 1, n. 3, p. 80-105, 2024.

SHARMA, A. Agentic AI security: threat modeling + controls for AI agents (permissions, tool-use constraints, auditability, kill-switch design). **SAMRIDDHI**, v. 17, n. 2, p. 47-60, 2025.

WANG, H.; POSKITT, C. M.; SUN, J. AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents. **arXiv**, 31 jul. 2025.

WILLIAMS, T.; LEE, J.; COSGROVE, J.; SAADE, T.; KANG, T. The infrastructure gap: why platform security cannot protect against agentic attacks. **Social Science Research Network**, 2025.

YANG, Y. Y.; MA, M.; HUANG, Y.; CHAI, H.; GONG, C.; GENG, H.; ZHOU, Y.; WEN, Y.; FANG, M.; CHEN, M.; GU, S.; JIN, M.; SPANOS, C.; YANG, Y.; ABBEEL, P.; SONG, D.; ZHANG, W.; WANG, J. Agentic Web: weaving the next web with AI agents. **arXiv**, 28 jul. 2025.