Além da Segurança do Modelo: Prontidão Institucional para a Adoção Segura de Inteligência Artificial no Setor Público Brasileiro

Alessandro Roosevelt Silva Ribeiro[1]
Maxwell Sarmento de Carvalho[1], **Daniela de Oliveira Moraes**[1]

[1]Universidade de Brasília (UnB)

alersr1010@gmail.com, maxwellsarmento@gmail.com

danni_danni@hotmail.com

Abstract. *Secure adoption of artificial intelligence in the public sector depends on organizational capabilities beyond model robustness. This short paper proposes an institutional readiness perspective and presents exploratory evidence from the 2024 iESGo open dataset, covering 387 Brazilian public organizations. Cybersecurity maturity and bureaucratic capacity are combined into a deterministic institutional vulnerability index, negatively associated with a results-orientation proxy ($\beta = -0.681$, $p < 0.001$, $R^2 = 0.456$). This association is exploratory and does not validate AI-security outcomes or technical vulnerability. The study instead examines institutional capabilities that may constitute antecedent conditions for the secure adoption of AI.*

Resumo. *A adoção segura de inteligência artificial no setor público depende de capacidades organizacionais que ultrapassam a robustez do modelo. Este artigo curto propõe uma perspectiva de prontidão institucional e apresenta evidências exploratórias dos dados abertos do iESGo 2024, referentes a 387 organizações públicas brasileiras. A maturidade cibernética e a capacidade burocrática são combinadas em um índice determinístico de vulnerabilidade institucional, negativamente associado a uma proxy de orientação para resultados ($\beta = -0{,}681$, $p < 0{,}001$, $R^2 = 0{,}456$). A associação é exploratória e não valida resultados de segurança de IA nem vulnerabilidade técnica. O estudo examina capacidades institucionais que podem constituir condições antecedentes à adoção segura de IA.*

1. Introdução

A adoção de modelos generativos, sistemas de apoio à decisão e agentes autônomos amplia a superfície de ataque dos serviços públicos digitais, incluindo riscos de envenenamento de dados, *prompt injection*, vazamento de contexto, extração de informações, comprometimento da cadeia de suprimentos e abuso de integrações [1, 3]. A análise restrita à robustez do modelo não captura as condições institucionais que determinam se esses riscos serão identificados, monitorados e tratados.

No setor público, falhas de IA podem afetar dados pessoais, continuidade de serviços e legitimidade decisória. Por isso, governança, gestão de riscos, controle de

acesso, resposta a incidentes, continuidade, auditoria e responsabilização constituem capacidades antecedentes: não garantem segurança, mas sua ausência torna a operação segura pouco plausível.

Este artigo propõe que a *prontidão institucional para IA segura* seja tratada como condição antecedente da segurança de sistemas de IA. O objetivo é responder a duas questões: **RQ1**: quais capacidades institucionais mensuradas em organizações públicas apresentam correspondência conceitual com funções de governança e segurança de IA? **RQ2**: como a insuficiência conjunta dessas capacidades se associa à orientação institucional para resultados? A RQ1 é examinada por mapeamento conceitual; a RQ2, por análise estatística exploratória. Assim, a contribuição combina uma ponte conceitual entre capacidades institucionais e controles de IA com uma operacionalização preliminar dessas capacidades no iESGo 2024.

2. Prontidão Institucional e Modelo de Ameaça

Definimos prontidão institucional para IA segura como a capacidade de uma organização para governar, proteger, monitorar e sustentar sistemas de IA ao longo de seu ciclo de vida. A noção complementa abordagens técnicas de segurança, pois controles de modelo não substituem processos organizacionais de decisão, resposta e responsabilização [1, 2].

O modelo considera como ativos dados de treinamento e inferência, documentos RAG, *prompts*, respostas, modelos, APIs, integrações, registros de auditoria e decisões apoiadas por IA. Os adversários incluem usuários externos, agentes internos, fornecedores comprometidos e atacantes da cadeia de suprimentos, capazes de submeter entradas maliciosas, manipular bases, explorar permissões, extrair informações ou abusar de ferramentas.

Para responder à RQ1, elaboramos mapeamento conceitual manual, não exaustivo, entre capacidades do iESGo e funções descritas no NIST AI RMF, NIST CSF e OWASP GenAI. O mapeamento admite correspondências múltiplas e não pressupõe equivalência normativa (Tabela 1).

Tabela 1. Capacidades institucionais e funções relacionadas à segurança de IA

Capacidade institucional	Função de IA relacionada
Gestão de riscos	Governança e gestão de riscos de IA [1]
Controle de acesso	Gestão de identidades, privilégios e acesso a ferramentas [2, 3]
Auditoria	Registro, monitoramento, rastreabilidade e responsabilização [1, 3]
Continuidade	Resiliência, recuperação e continuidade operacional [2]
Resposta a incidentes	Detecção, resposta e recuperação diante de eventos adversos [2, 3]

A correspondência responde à RQ1 no nível de fundamento institucional; não demonstra implementação de controles específicos nem resistência a ataques sobre modelos, RAG ou agentes.

3. Método

Foram utilizados os dados abertos do iESGo 2024, do Tribunal de Contas da União, com 387 organizações públicas [4]. A maturidade cibernética (MC) reuniu 53 itens sobre riscos de TI, segurança, continuidade, incidentes, acesso e desenvolvimento seguro; a capacidade burocrática (CB), 75 itens sobre governança, riscos, estratégia, processos, pessoas e monitoramento; e o resultado alcançado (RA), 15 itens sobre metas, resultados, serviços digitais e transparência.

As respostas foram normalizadas em $[0, 1]$ e cada escore corresponde à média dos itens válidos, com cobertura mínima de 70%; todas as organizações apresentaram cobertura integral. A vulnerabilidade institucional agrega as insuficiências de MC e CB por ponderação linear [5]:

$$\Omega_i = \lambda(1 - MC_i) + (1 - \lambda)(1 - CB_i), \quad 0 \leq \lambda \leq 1. \tag{1}$$

O parâmetro λ pondera a insuficiência de MC e $(1 - \lambda)$ preserva a soma dos pesos em 1 e o índice em $[0, 1]$. A especificação principal fixa $\lambda = 0{,}5$ *ex ante*, com pesos iguais e sem calibração pelo desfecho; 0,25 e 0,75 são usados apenas na análise de sensibilidade.

A associação com RA foi estimada por regressão linear via mínimos quadrados ordinários, com erros-padrão robustos HC3. Como Ω é combinação determinística de MC e CB, a especificação alternativa estima os componentes separadamente:

$$RA_i = \alpha + \beta\Omega_i + \varepsilon_i, \tag{2}$$

$$RA_i = \alpha + \beta_1 MC_i + \beta_2 CB_i + \varepsilon_i. \tag{3}$$

Os modelos são associativos, não causais, pois os dados são transversais e autodeclarados.

4. Resultados Exploratórios

A Tabela 2 resume os indicadores. A média de MC foi inferior à média de CB, sugerindo que práticas gerais de governança e planejamento estão mais desenvolvidas do que práticas especificamente cibernéticas no conjunto examinado.

Tabela 2. Estatísticas e resultados principais

Indicador ou modelo	Média	Desvio-padrão	Resultado
MC	0,468	0,299	53 itens
CB	0,617	0,284	75 itens
RA	0,578	0,264	15 itens
Ω	0,458	0,262	$\lambda = 0{,}5$
$RA \sim \Omega$	–	–	$\beta = -0{,}681$, $R^2 = 0{,}456$
$RA \sim MC + CB$	–	–	$R^2 = 0{,}470$

Na regressão linear, Ω apresentou coeficiente negativo ($\beta = -0{,}681$, $p < 0{,}001$): aumento de 0,10 no índice associa-se a uma redução média de 0,068 em RA. No modelo decomposto, MC ($\beta = 0{,}223$, $p < 0{,}001$) e CB ($\beta = 0{,}466$, $p < 0{,}001$) mantiveram associações positivas; o R^2 passou de 0,456 para 0,470.

O sinal negativo foi preservado para $\lambda \in \{0{,}25; 0{,}50; 0{,}75\}$. A forma aditiva, porém, permite compensação entre dimensões; estudos futuros devem compará-la a alternativas não compensatórias, como $1 - \sqrt{MC \cdot CB}$ e $1 - \min(MC, CB)$.

5. Discussão e Limitações

Os resultados respondem às questões em níveis distintos. Para a RQ1, o mapeamento identifica correspondências conceituais entre capacidades institucionais e funções de governança e segurança de IA. Para a RQ2, maior insuficiência conjunta de MC e CB associa-se a menor orientação para resultados. A evidência apoia a relevância de capacidades antecedentes, mas não valida Ω como medida de vulnerabilidade técnica ou prontidão específica para IA.

As limitações são substantivas: os dados não contêm itens específicos de IA, MLOps, LLMOps, *red teaming*, RAG ou *drift*; preditores e desfecho provêm do mesmo questionário autodeclarado, com risco de viés de método comum [6]; RA não mede desfechos de segurança; e o desenho transversal impede inferência causal. Permanece em aberto a validação do índice contra inventários reais de IA, incidentes, auditorias, testes adversariais e métricas de detecção e resposta.

6. Conclusão

A principal contribuição é estruturar a prontidão institucional como camada antecedente da segurança de IA e relacionar capacidades já mensuradas no setor público a funções de controle de IA. A análise de 387 organizações oferece operacionalização preliminar dessa perspectiva: Ω associa-se à orientação para resultados, mas não mede vulnerabilidade técnica nem resistência a ataques adversariais. A continuidade da pesquisa requer validação externa, longitudinal e adversarial com sistemas de IA observados em produção.

Referências

[1] National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, Gaithersburg, MD.

[2] National Institute of Standards and Technology (2024). *The NIST Cybersecurity Framework (CSF) 2.0*. NIST CSWP 29, Gaithersburg, MD.

[3] OWASP Foundation (2025). *OWASP Top 10 for Large Language Model Applications*. OWASP GenAI Security Project.

[4] Tribunal de Contas da União (2024). *iESGo 2024: Índice ESG de organizações públicas*. Dados abertos e documentação do levantamento, Brasília, DF.

[5] OECD and Joint Research Centre (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris.

[6] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5):879–903.