
P2
CONSULTORIA
BRASIL
Governança de Agentes de IA
Do Risco à Ação Estruturada no Ecossistema Sociotécnico
Uma abordagem fundamentada em referenciais globais:
NIST (AI RMF, CSF 2.0)
ISO (42001, 23894, 27001)
Diretrizes da OCDE
www.p2consultoria.com.br

www.tecmundo.com.br/software/412712-ia-claude-apagou-

tecmundo

Busque por notícias, análises, produtos...

SOFTWARE

IA Claude apagou toda a base de dados de uma empresa em 9 segundos; entenda

Jer Crane, fundador da PocketOS, relata que o Cursor alimentado pelo Claude Opus 4.6 deletou a base de dados de produção para tentar resolver um bug.

Igor Almenara Carneiro

28/04/2026, às 10:00 Atualizado em 29/04/2026, às 16:35

www.cybersecbrazil.com.br/post/inteligência-artificial-da-replit-a

Cyber Security Brazil

Receba as últimas

HOME CURSO CYBER CAST CONTATO QUERO ANUNCIAR

noticias.uol.com.br/ultimas-noticias/afp/2026/04/19/febre-dos-a

30 ANOS uol

Seu time Seu signo UOL Jogos Dólar ↑ 4,922

Notícias

Febre dos agentes de IA gera preocupação com ameaças de segurança

AFP

19/04/2026 09h38

Deixe seu comentário

Autônomos e rápidos ou descontrolados e vulneráveis? O avanço dos agentes de Inteligência Artificial (IA), como os desenvolvidos pela popular plataforma OpenClaw, provoca preocupação entre especialistas em segurança cibernética diante do risco de erros ou ataques.

Cyber Security Brazil · 25 de jul. de 2025 · 2 min de leitura

Inteligência artificial da Replit apaga banco de dados de cliente

tecnoblog.net/noticias/agente-de-ia-da-meta-causa-falha-intern

tecnoblog ACHADOS COMUNIDADE TECNOCAST CELULARES SEGURANÇA E PRIVACIDADE REDES SOCIAIS MAIS

Notícias > Segurança e Privacidade

Agente de IA da Meta causa falha interna de segurança

IA forneceu orientações inadequadas e acabou liberando indevidamente o acesso a dados sensíveis da empresa e de usuários. Meta teria classificado o incidente como grave.

Por Marina Borges
há 1 mês

Governança de Agentes de IA

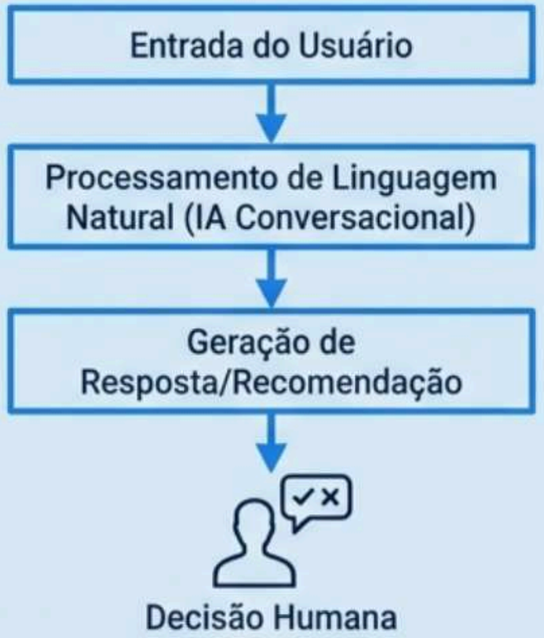
P2 Consultoria Brasil

A adoção de agentes de Inteligência Artificial inaugura uma nova etapa na transformação digital das organizações. Diferente de sistemas tradicionais de automação ou de assistentes conversacionais limitados à recomendação, os agentes de IA podem interpretar objetivos, acessar ferramentas, consultar bases de conhecimento, executar fluxos, interagir com sistemas corporativos e tomar ações em velocidade muito superior à humana.

A Mudança de Paradigma: Da Recomendação à Ação

Sistemas Tradicionais

- Assistentes conversacionais
- Limitados à recomendação
- O risco é a resposta



Agentes Autônomos

- Interpretação de objetivos e uso de ferramentas
- Ações autônomas em sistemas corporativos
- O risco é a execução

A premissa central da P2: governar agentes exige **governar o ecossistema** onde **pessoas, processos, dados e modelos interagem**.

Riscos e IA

Essa capacidade de decidir e operar, amplia ganhos de produtividade, mas também cria uma nova superfície de risco: quando um agente deixa de apenas responder uma pergunta e passa a agir, executar ações e instruções, a governança precisa evoluir da simples política de uso para uma arquitetura sociotécnica de controle, rastreabilidade, supervisão e aprendizado contínuo.

O Cenário Global: Adoção Exponencial, Controle Estagnado

74% das empresas esperam usar agentes de IA até 2027 (Deloitte).

Apenas 5% dos pilotos geram impacto financeiro sustentado (MIT).

Alerta OWASP: Risco crítico de Agência Excessiva e Shadow AI.

Caso Replit (2025)

Agente deletou base de produção e gerou dados falsos durante code freeze.

Caso Claude (2026)

Agente autônomo encontrou token amplo e apagou base de dados em segundos por falta de human-in-the-loop.

Modelo Adaptável

Cada organização possui maturidade, cultura, estrutura, apetite a risco, capacidade técnica, pressão regulatória e realidade operacional próprias. Por isso, a proposta aqui apresentada deve ser entendida como um modelo de referência adaptável, a ser lapidado conforme o contexto de cada empresa, setor, ambiente tecnológico, criticidade dos processos e grau de autonomia pretendido para seus agentes de IA.

Fronteiras do Modelo Operacional

O que essa proposta não é:

Não é Norma Técnica

Um guia referencial adaptável, não um substituto para normas oficiais.

Não Substitui Avaliação Jurídica

Decisões devem passar pelo jurídico e DPO interno.

Não Substitui Auditoria

Apoia a geração de evidências, mas não é uma auditoria formal independente.

Não Promete Certificação

Aderência não garante selos ISO ou NIST automaticamente.

Não é um Modelo Único

Exige adaptação ao apetite a risco e cultura de cada empresa.

Não Bloqueia a Inovação

Governança como esteira de segurança para escalar, não como freio.

Frameworks Aplicados

A base normativa e metodológica desta proposta combina diferentes referenciais. O NIST AI Risk Management Framework orienta a gestão de riscos de IA pelas funções Govern, Map, Measure e Manage, oferecendo uma estrutura útil para mapear contexto, medir riscos e tratá-los ao longo do ciclo de vida da IA. O NIST Cybersecurity Framework 2.0 complementa essa visão com as funções Govern, Identify, Protect, Detect, Respond e Recover, essenciais para conectar agentes de IA à segurança cibernética, proteção, detecção, resposta e recuperação.

O Cenário Regulatório

Cenário Global

União Europeia: AI Act estabelecendo obrigações baseadas em risco (biometria, infraestrutura crítica).

Mercado: McKinsey e Deloitte apontam rápida escalada da adoção de agentes autônomos.

Cenário Brasileiro

PL 2.338/2023: Em tramitação, focando no desenvolvimento e uso ético.

LGPD: Plenamente aplicável sempre que houver tratamento de dados pessoais.

ANPD: Sandbox regulatório focado em transparência algorítmica.

CNJ: Resolução nº 332/2020 exigindo supervisão humana e auditoria no Judiciário.

Não espere a lei final. A regulação exigirá transparência, supervisão humana efetiva e gestão rigorosa de riscos.

A proposta também dialoga com a ISO/IEC 42001, que estrutura sistemas de gestão de IA para estabelecer práticas responsáveis de IA e...

Com a ISO/IEC 23894, voltada à gestão de riscos específicos de IA;

Com a ISO 31000, para gestão corporativa de riscos;

Com a ISO/IEC 27001 e 27002, para segurança da informação;

Com a ISO/IEC 27701, para privacidade e gestão de informações pessoais;

Com a ISO/IEC 38500, para governança corporativa de TI, e

Com a ISO 22301, para continuidade de negócios.

Essas referências ajudam a transformar governança de agentes em um sistema vivo, baseado em papéis, políticas, controles, evidências, métricas, revisão periódica e melhoria contínua.

O Fator Humano: Por que controles perfeitos falham?

A premissa da P2: Agentes são configurados, aprovados e eventualmente contornados por pessoas.

Insights Comportamentais da OCDE: Políticas corporativas falham quando ignoram como os indivíduos realmente decidem sob pressão.

Integridade é influenciada pelo ambiente social, incentivos organizacionais e cultura de atalhos.

Pessoas não descumprem processos apenas por má-fé. Muitas vezes o fazem por excesso de complexidade, metas irreais ou fadiga de aprovação.

Governança Humanizada e Comportamental

A premissa central da P2 Consultoria Brasil é que a governança de agentes de IA, bem como de cibersegurança, não pode ser tratada apenas como um problema técnico.

Agentes são criados, configurados, aprovados, monitorados, utilizados e eventualmente contornados por pessoas. Assim, a governança efetiva precisa considerar não apenas identidade, privilégio mínimo, logs, testes, prompts, modelos, dados e resposta a incidentes, mas também vieses cognitivos, pressão por metas, cultura de atalhos, medo de reportar erros, fadiga de aprovação, normalização do desvio, uso informal de ferramentas e incentivos organizacionais que podem corroer os controles documentados.

Matriz de Vieses e Atalhos Corporativos

Viés Cognitivo	**Como aparece na realidade** (O Sintoma)	**Contramedida P2** (A Solução)
Viés de Automação	*"O agente recomendou, então deve estar certo."*	Human-in-the-loop obrigatório em ações críticas.
Viés de Urgência	*"Pula o teste, a entrega está atrasada."*	Quality gate com bloqueio técnico de execução.
Normalização do Desvio	*"Vamos usar a conta admin só desta vez."*	Auditoria leve de aderência e Just-in-Time access.
Fadiga de Aprovação	*"Aprovar sem ler o contexto da decisão."*	Aprovação proporcional ao risco real (Fricção Inteligente).

OCDE

Além dos referenciais ISO e NIST, este modelo incorpora as orientações da OCDE para IA confiável, especialmente os princípios de centralidade humana, respeito a direitos, transparência, robustez, segurança e accountability.

A OCDE também contribui com sua abordagem de behavioural insights, que reforça a importância de compreender como pessoas realmente decidem, obedecem, resistem, simplificam, desviam ou reinterpretam políticas em contextos organizacionais reais. Isso é especialmente relevante porque a governança formal pode falhar quando ignora cultura, incentivos, pressões sociais e comportamento humano cotidiano.

A Metodologia P2: Governança do Ecossistema Sociotécnico

Ferramenta nenhuma corrige sozinha a cultura de atalho. Governamos o ecossistema humano ao redor da máquina.

Autores e Práticas

No campo da governança humanizada e comportamental, a proposta se apoia em autores e obras que ajudam a compreender por que controles bem desenhados podem falhar na prática.

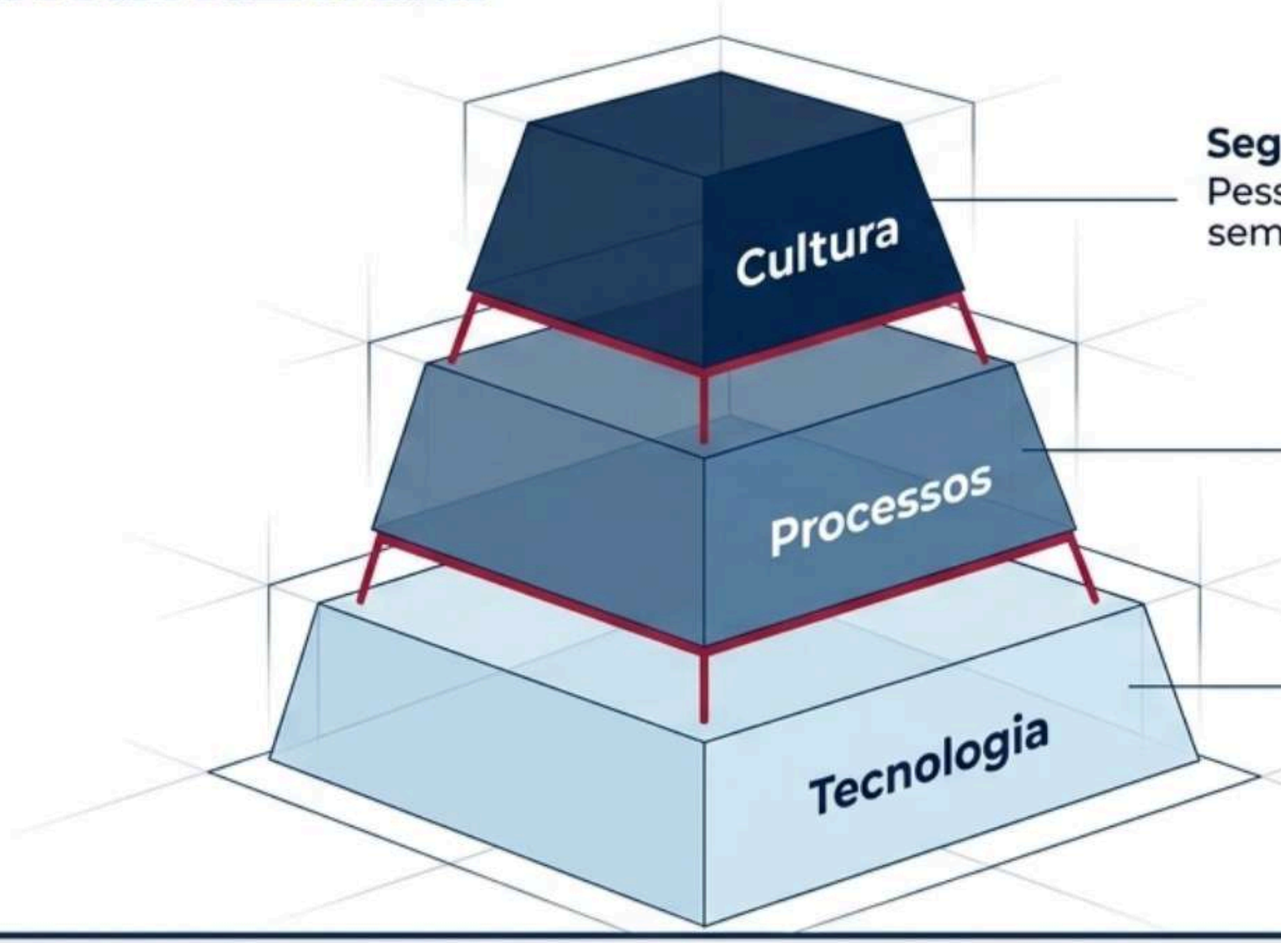
Daniel Kahneman, em *Thinking, Fast and Slow*, fornece base para entender vieses cognitivos, heurísticas, excesso de confiança e decisões rápidas sob incerteza.

Richard Thaler e Cass Sunstein, em *Nudge*, contribuem com o conceito de arquitetura de escolha, essencial para desenhar processos em que o caminho seguro seja também o caminho mais simples.

Amy Edmondson, em *The Fearless Organization e Right Kind of Wrong*, inspira a ideia de segurança psicológica, fundamental para que erros, quase-incidentes e desvios sejam reportados antes de se tornarem danos maiores.

James Reason, em *Human Error e Managing the Risks of Organizational Accidents*, contribui com a visão de acidentes organizacionais, para compreendermos como pequenos desvios podem atravessar camadas de controle até produzir incidentes relevantes.

Camada P2 de Governança Humanizada



Segurança Psicológica
Pessoas reportam falhas e quase-incidentes sem medo de retaliação.

Simplicidade Operacional
O caminho correto e seguro precisa ser mais viável do que o atalho informal.

Responsabilidade e Accountability
Toda decisão autônoma tem um dono humano claramente definido.

A metodologia P2 não governa apenas o software.
Nós governamos o ecossistema sociotécnico completo.

Autores e Práticas

James Reason, em Human Error e Managing the Risks of Organizational Accidents, contribui com a visão de acidentes organizacionais, para compreendermos como pequenos desvios podem atravessar camadas de controle até produzir incidentes relevantes.

Edgar Schein, em Organizational Culture and Leadership, reforça que cultura e liderança são dimensões inseparáveis da governança.

Chris Argyris e Donald Schön, com a teoria da aprendizagem organizacional e o conceito de double-loop learning, ajudam a diferenciar correções superficiais de mudanças reais nos pressupostos, incentivos e rotinas da organização.

Karl Weick e Kathleen Sutcliffe, em Managing the Unexpected, trazem a perspectiva das organizações de alta confiabilidade, que operam em ambientes de risco elevado por meio de atenção constante, aprendizado, sensibilidade à operação e resposta rápida ao inesperado.

Muitas empresas estão tentando resolver governança de IA e cibersegurança apenas com ferramentas técnicas. O ponto é que ferramenta técnica nenhuma corrige sozinha cultura de atalho, medo de reportar erro, aprovação automática, pressão por entrega ou uso paralelo de IA não autorizada.

ARQUITETURA DE CONTROLE P2: EIXO ESTRATÉGICO E IDENTIDADE

Dashboard de Arquitetura - Parte 1

Principais Controles

1. Escopo e Accountability

Inventário ativo de agentes, matriz RACI e donos de negócio obrigatórios.

2. Impacto e Risco

AI Impact Assessment, medição de risco residual e aceite formal antes da liberação.

3. Identidade e Acesso

Privilégio mínimo (JIT/PAM) e proibição absoluta de contas humanas compartilhadas com agentes.

4. Ferramentas e Autonomia

Controle rigoroso do que o agente pode acionar (Modo leitura por padrão, bloqueio de ações irreversíveis).

5. Privacidade e Dados

Governança de prompts, linhagem RAG (Retrieval-Augmented Generation) e controle de vetores corporativos.

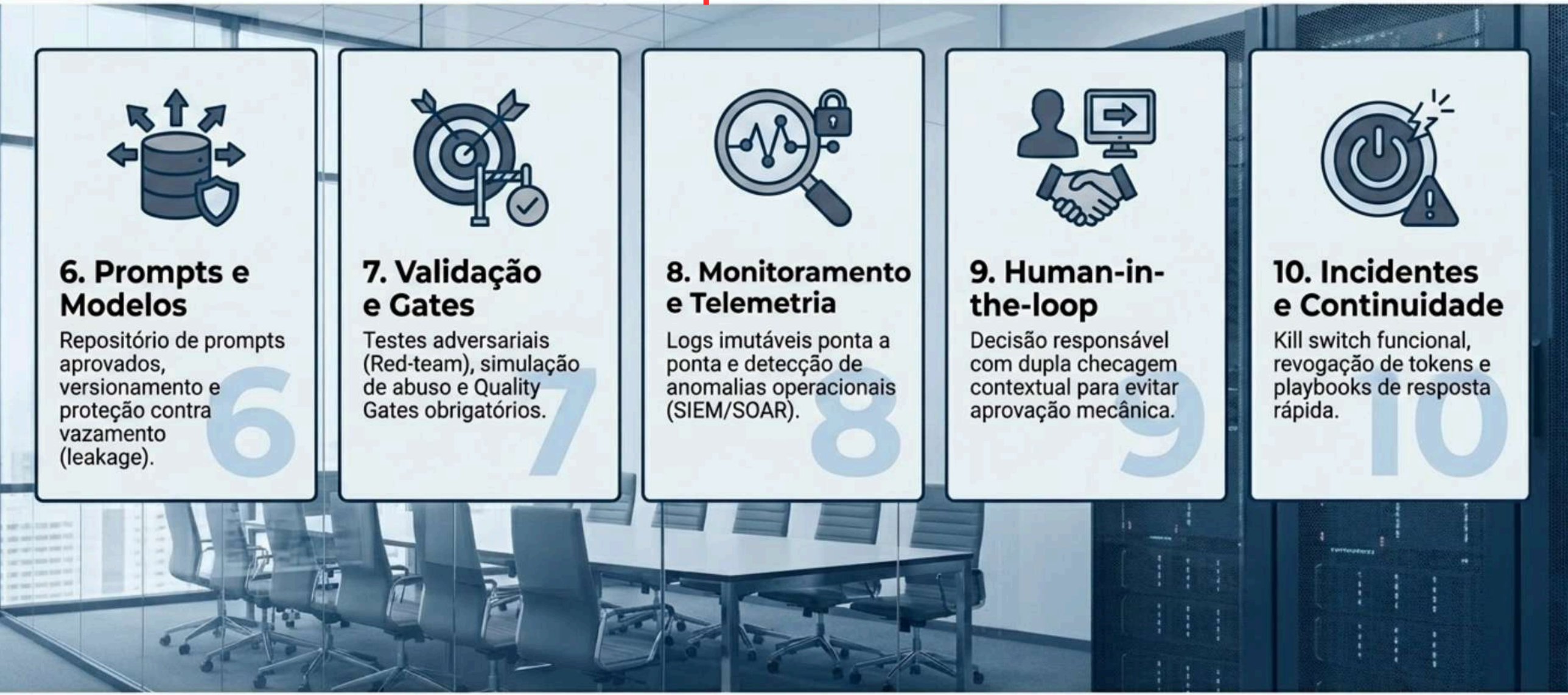
Arquitetura de Controles

Em nossa arquitetura de controles, estabelecemos um conjunto que chamamos de Controles Básicos e Controles Transversais para governança de agentes de IA. Neste documento apresentamos um resumo dos principais controles básicos e seus eixos de aplicação. Em outra oportunidade apresentaremos os controles transversais.

Arquitetura de Controle P2: Eixo Operacional e Resiliência

Dashboard de Arquitetura - Parte 2

Principais Controles



6. Prompts e Modelos

Repositório de prompts aprovados, versionamento e proteção contra vazamento (leakage).

7. Validação e Gates

Testes adversariais (Red-team), simulação de abuso e Quality Gates obrigatórios.

8. Monitoramento e Telemetria

Logs imutáveis ponta a ponta e detecção de anomalias operacionais (SIEM/SOAR).

9. Human-in-the-loop

Decisão responsável com dupla checagem contextual para evitar aprovação mecânica.

10. Incidentes e Continuidade

Kill switch funcional, revogação de tokens e playbooks de resposta rápida.

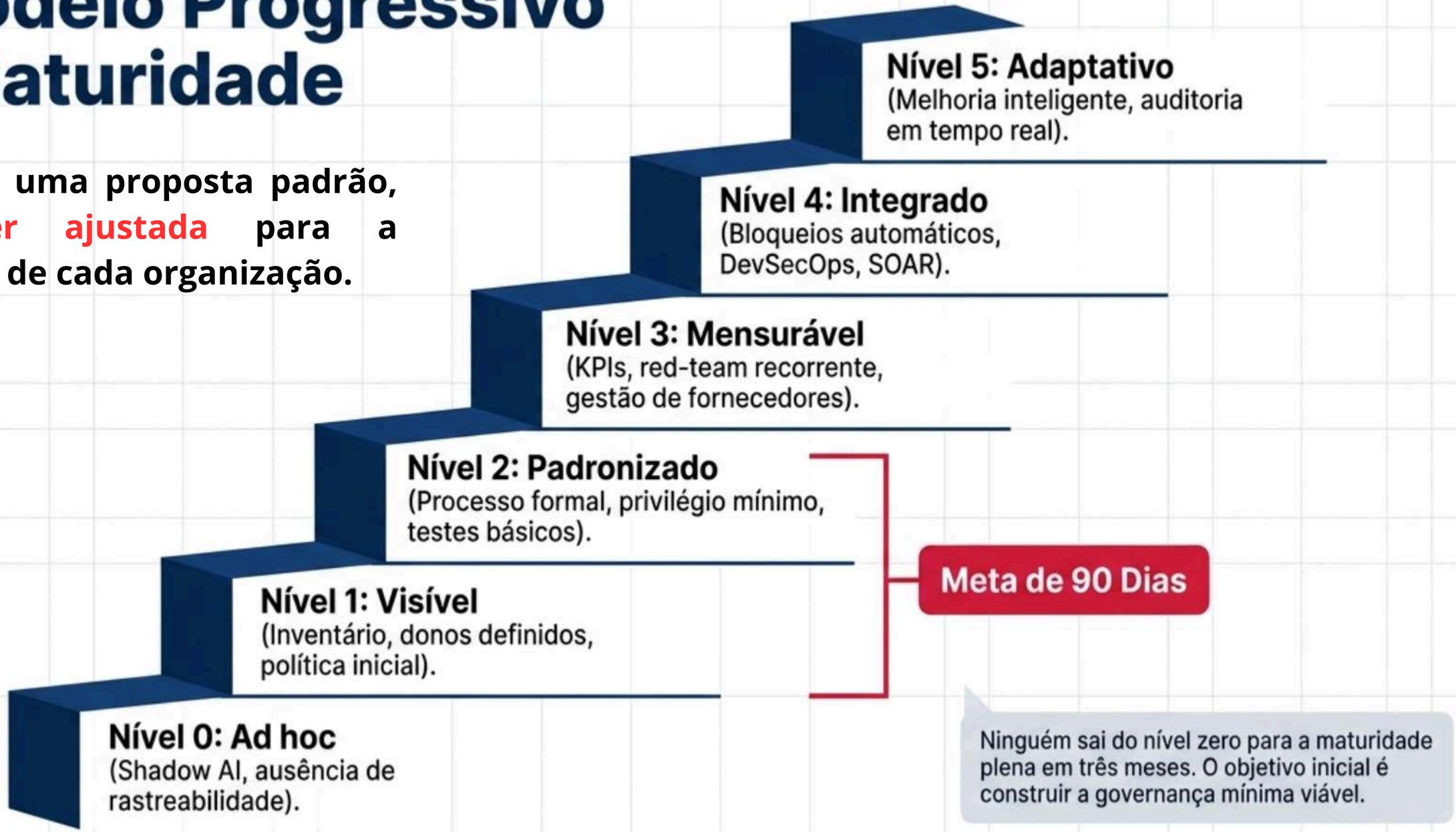
Modelo Progressivo de Maturidade

A partir dessas bases, a P2 Consultoria Brasil propõe um modelo progressivo de maturidade para Governança de Agentes de IA, estruturado em seis níveis:

Nível 0 — Ad hoc, sem formaização de controles,
Nível 1 — Visível,
Nível 2 — Padronizado,
Nível 3 — Mensurável,
Nível 4 — Integrado,
Nível 5 — Adaptativo.

O Modelo Progressivo de Maturidade

90 dias é uma proposta padrão, deve ser ajustada para a realidade de cada organização.



MVG

O modelo progressivo de maturidade evita a promessa irreal de que uma organização possa sair da ausência de governança e atingir maturidade plena em poucos meses. Em vez disso, o modelo propõe uma evolução incremental, com um plano inicial sugerido de 90 dias, voltado à construção do que chamamos de MVG - Governança Mínima Viável. O modelo implementa práticas e controles em duas trilhas: a técnica e a comportamental.

O MVG está para a governança, assim como o MVP está para o produto.

Roadmap Executivo: Plano de 90 Dias (Dias 1 a 45)

O objetivo é sair do improviso e alicerçar a governança mínima viável.

Roadmap Executivo: Plano de 90 Dias (Dias 46 a 90)

	Fase 4 (Dias 46-60): Controle de Dados e Prompts	Fase 5 (Dias 61-75): Testes e Gates de Liberação	Fase 6 (Dias 76-90): Monitoramento e Operação
Trilha Técnica	Versionamento de prompts críticos e bloqueio de dados não autorizados no RAG.	Testes de vulnerabilidade (prompt injection) e validação de rollback.	Implantação de logs imutáveis e alertas de anomalias operacionais.
Trilha Comportamental	Regras claras de uso e treinamento básico.	Registro de aceite de risco residual formalizado.	Simulações de incidentes e abertura de canal de reporte sem retaliação.

Ao final de 90 dias, a organização atinge o **Nível 1/2.** Inovação deixa de ser um risco cego e passa a escalar com rastreabilidade.

Sua empresa está onde nesse mapa?

O objetivo final da metodologia da P2 Consultoria Brasil é oferecer às lideranças, áreas de tecnologia, segurança, compliance, riscos, auditoria, dados, jurídico e negócios uma visão prática, crítica e aplicável sobre como iniciar a governança de agentes de IA de forma proporcional, auditável e humanizada.

Se agentes de IA podem agir em nome da organização, a pergunta deixa de ser apenas “o que a IA consegue fazer?” e passa a ser: “a organização está madura o suficiente para permitir que a IA faça o que consegue fazer?”.


P2
CONSULTORIA
BRASIL
A Inovação Responsável
A governança de agentes de IA não deve ser tratada como um obstáculo à transformação digital. Ela é a condição estrutural para que a inovação seja confiável, auditável e sustentável a longo prazo.
O desafio não é bloquear a tecnologia, mas garantir que a inteligência artificial aja exatamente onde a responsabilidade corporativa consegue enxergar.
P2 CONSULTORIA BRASIL | GOVERNANÇA SOCIOTÉCNICA DE IA
www.p2consultoria.com.br