



sandbox
regulatório

Metodologia de Testagem

Projeto-Piloto de *Sandbox* Regulatório em
Inteligência Artificial e Proteção de Dados

ANPD – Agência Nacional de Proteção de Dados

Diretor-Presidente

Waldemar Gonçalves Ortunho Junior

Diretores

Iagê Zendron Miola

Lorena Giuberti Coutinho

Miriam Wimmer

Comissão de Sandbox

Diego André Sant'Ana

Fábio Silveira Vidal

Lucas Costa dos Anjos

Maria Carolina Ferreira da Silva

Marcus Vinícius Rossi da Rocha

Jeferson Dias Barbosa

Equipe de Supervisão

Albert França Josué Costa

Uberson Rossa

Equipe de Apoio Administrativo

Gabriel Carvalho Moretto

Marcia Valéria Paixão

Raphael Gonçalves Vanderlei

ANPD

SCN, Qd. 6, Conj. A, Ed. Venâncio 3000, Bl. A, 9º andar
Brasília, DF Brasil 70716-900
t. (61) 2025-8101 www.gov.br/anpd

**Centro de Inteligência Artificial e Aprendizado de Máquina Universidade de São Paulo
-CIAAM/USP**

Cristina Godoy Bernardo de Oliveira (Coordenadora)

Anarosa Brandão (Responsável pela Equipe Técnica)

Juliano Maranhão (Responsável pela Equipe Técnica)

Carlos Portugal (Responsável pela Equipe Técnica)

João Paulo Cândia Veiga (Responsável pela Equipe Técnica)

Bruno Albertini (Pesquisador – Área Técnica)

Fernanda Valle Versiani (Pesquisadora – Área Regulatória)

Luciano Gallegos (Pesquisador – Área Técnica)

Paloma Saldanha (Pesquisadora – Área Regulatória)

Rafael Meira Silva (Pesquisador – Área Regulatória)

Giovana Carneiro (Pesquisadora – Área Regulatória)

Thiago Marcílio (Pesquisador – Área Regulatória)

Anthony Novaes (Pesquisador – Área Regulatória)

Consultores Internacionais

João Pita Costa (IRCAI UNESCO)

Vasilka Sancin (Universidade de Ljubljana e Juíza da Corte Europeia de Direitos Humanos)

Colaboradores HARPIA

Sarajane Peres (Pesquisadora Responsável USP)

Vinícius B. Matos (Pesquisador)

Ricardo Grava (Pesquisador)

Colaboradores Médicos (Testes de LLM na Área de Saúde)

Daniele Thais de Almeida Santos

Paulo Victor Gomes Ferreira

Haroldo Rodrigues

Matheus Ananias

Giuliano Bongiovanni

SUMÁRIO

1.Introdução	5
2. Proposição de metodologias para supervisão e acompanhamento dos testes dos participantes	6
2.1 Definição de indicadores e variáveis para avaliação de resultados e impactos	6
2.2 Desenho dos métodos de coleta e análise de dados referentes à testagem.....	10
2.2.1 Fase A - Preparação e Detalhamento	11
2.2.2 Fase B - Experimentação e Gestão de Riscos.....	13
2.2.3 Fase C - Avaliação e Ajuste	14
2.3 Consolidação dos dados e estruturação do plano de monitoramento e avaliação	16
3 Plano de <i>Sandbox</i> e Plano de Descontinuidade	20
3.1 Plano de Sandbox (PdS)	20
3.2 Plano de Descontinuidade (PdD).....	22

1.Introdução

O projeto-piloto de *Sandbox* Regulatório em Inteligência Artificial e Proteção de Dados da Agência Nacional de Proteção de Dados (ANPD) constitui um ambiente regulatório experimental, controlado e supervisionado, voltado à testagem de projetos inovadores que utilizam sistemas de inteligência artificial e envolvem tratamento de dados pessoais. A iniciativa busca promover a inovação responsável, a transparência algorítmica e a conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD), permitindo que soluções baseadas em inteligência artificial sejam acompanhadas em contexto estruturado de orientação, monitoramento e avaliação.

O objetivo deste documento é apresentar a metodologia desenvolvida pela equipe do Centro de Inteligência Artificial e Aprendizado de Máquina da Universidade de São Paulo (CIAAM/USP) e validada pela Comissão de *Sandbox* da ANPD, empregada na supervisão e no acompanhamento dos testes dos participantes, identificando os métodos e as técnicas necessários para viabilizar o alcance do escopo do projeto, qual seja: avaliar pontos relacionados à transparência, à explicabilidade e à proteção de dados pessoais em soluções inovadoras.

As atividades de testagem estão estruturadas em seis ciclos compostos por três fases: (i) Fase A - Pré-teste; (ii) Fase B - Teste; e (iii) Fase C - Pós-teste. Na fase A, os participantes preenchem fichas a respeito de suas aplicações, metodologias e estágio de desenvolvimento de módulos ou componentes. Na fase B, testes específicos para cada participante são executados, considerando as informações coletadas na fase A e os artefatos de software fornecidos pelos participantes. Ao final de cada ciclo, na fase C, pareceres são apresentados individualmente para cada participante. Tais pareceres orientam o próximo ciclo de testagem, visando melhorias técnicas, éticas e regulatórias.

Adicionalmente, será dedicado um tópico à apresentação dos indicadores e das variáveis utilizados para a avaliação dos resultados e dos impactos do *sandbox* regulatório. Esses elementos são essenciais para analisar as informações coletadas e acompanhar, de forma sistemática, os resultados e os impactos observados ao longo de cada ciclo de testagem. Com base nessa avaliação, será possível identificar oportunidades de aprimoramento e promover ajustes contínuos, contribuindo para a evolução das atividades desenvolvidas em cada ciclo.

2. Proposição de metodologias para supervisão e acompanhamento dos testes dos participantes

A metodologia de testagem e supervisão do *Sandbox* foi estruturada em seis ciclos. Com exceção do Ciclo 1, cada ciclo é subdividido em três fases: Fase A (pré-testagem), Fase B (testagem) e Fase C (pós-testagem), conforme Figura 1.



Figura 1: Ciclos de Testagem do Sandbox em Inteligência Artificial e Proteção de Dados.

O Ciclo 1 possui estrutura distinta, em razão da necessidade de adaptação do ambiente de testes do projeto e da coleta das informações iniciais dos agentes participantes. Nos ciclos 2 a 6, as atividades são organizadas nas três fases mencionadas anteriormente.

Por fim, vale ressaltar que os escopos dos ciclos de testagem foram definidos pelos participantes no Plano de *Sandbox*¹ que foi aprovado na fase de treinamento².

2.1 Definição de indicadores e variáveis para avaliação de resultados e impactos

A avaliação dos resultados do *Sandbox* Regulatório de IA e Proteção de Dados deve ser estruturada com base em indicadores e variáveis. Esses instrumentos devem permitir o acompanhamento, de forma cumulativa e comparável entre os ciclos de testagem, da evolução técnica, regulatória, ética e organizacional das soluções avaliadas.

¹ Mais informações sobre o Plano de *Sandbox* estão disponíveis em: <https://sites.usp.br/sandbox/plano-de-sandbox/>.

² Mais informações estão disponíveis em: <https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-conclui-fase-de-nivelamento-do-projeto-sandbox>

A definição de indicadores não deve se limitar à verificação do desempenho técnico dos modelos, mas também deve abranger a capacidade da solução de demonstrar aderência ao uso esperado, controle de riscos, proteção de titulares de dados pessoais, transparência, explicabilidade, governança, maturidade tecnológica e aprendizagem regulatória.

Entende-se por variável o dado observável coletado ao longo dos ciclos de testagem, como setor de aplicação, tipo de solução, arquitetura do modelo, categorias de dados tratados, existência de intervenção humana, nível de maturidade tecnológica, riscos identificados, medidas de mitigação, evidências documentais e manifestações de *stakeholders*.

O indicador é a medida construída a partir dessas variáveis para permitir a avaliação objetiva ou qualitativamente estruturada dos resultados, como variação do nível de risco residual, evolução da maturidade tecnológica, percentual de critérios éticos atendidos, grau de completude documental, taxa de implementação de medidas de mitigação, aderência aos usos esperados e grau de aprendizagem regulatória produzido pelo ciclo de testagem.

Para fins de avaliação, a unidade de análise será organizada em cinco níveis complementares:

- solução de inteligência artificial testada;
- componente tecnológico crítico, quando aplicável;
- ciclo de testagem;
- risco ou requisito regulatório avaliado; e
- medida de mitigação ou aprimoramento correspondente.

Essa estrutura permite que a avaliação de impacto seja longitudinal, isto é, baseada na comparação entre a situação inicial, as alterações observadas em cada ciclo e os resultados consolidados ao final da testagem. Sempre que possível, os indicadores deverão registrar linha de base, valor observado por ciclo, evidência correspondente, responsável pela informação, limitação metodológica e interpretação regulatória.

A matriz de indicadores deve contemplar, no mínimo, seis dimensões: i) técnico-funcional; ii) de dados, privacidade e segurança; iii) de maturidade tecnológica; iv) de realismo experimental e benefícios esperados; v) de riscos, mitigação e impacto residual; e vi) de governança e aprendizado regulatório (Figura 2).

	Dimensões	Nível 1  Produtos imediatos	Nível 2  Resultados intermediários	Nível 3  Impacto regulatório e institucional
	1. Técnico-funcional	- Ficha técnica - Arquitetura	- Estabilização da solução - Ajustes de escopo	- Critérios mínimos para descrição de sistemas
	2. Dados, privacidade e segurança	- Mapeamento de dados - Fluxos - Controles e logs	- Melhoria de privacidade - Reforço de segurança	- Parâmetros para proteção de dados e segurança
	3. Maturidade tecnológica	- Matriz TRL - ETCs identificados	- Evolução do TRL - Validação em ambiente relevante	- Referências para maturidade tecnológica
	4. Realismo experimental e benefícios esperados	- Plano de testes - Testes executados - Evidências	- Validação de benefícios - Revisão de premissas	- Condições para uso, continuidade ou restrição
	5. Riscos, mitigação e impacto residual	- Matriz de riscos - Plano de mitigação	- Redução do risco residual - Medidas implementadas	- Parâmetros para classificação e tratamento de riscos
	6. Governança, transparência e aprendizagem regulatória	- Papéis e responsáveis - Documentação - Explicabilidade	- Melhoria da governança - Mais transparência e accountability	- Aprendizados replicáveis - Critérios para supervisão futura



A avaliação acompanha o que foi produzido, o que mudou durante a testagem e o que pode orientar futuras decisões regulatórias.

Figura 2: Indicadores de avaliação dos resultados e impactos.

As seis dimensões estão detalhadas a seguir:

I. Dimensão técnico-funcional

A dimensão técnico-funcional é dedicada à descrição da arquitetura da solução, dos inputs relevantes, do pipeline de inferência, das APIs utilizadas, dos cenários de uso e das limitações de escopo. Seus indicadores devem medir, por exemplo, a estabilidade da arquitetura entre ciclos, a rastreabilidade dos inputs, a adequação do modelo ao uso declarado, a ocorrência de alterações relevantes de premissas e a aderência da solução aos limites operacionais definidos no *sandbox*.

II. Dimensão de dados, privacidade e segurança

A dimensão de dados, privacidade e segurança deve acompanhar os fluxos de dados, as categorias de dados pessoais e sensíveis, os métodos de anonimização ou pseudonimização, os controles de acesso, os mecanismos de backup, a guarda de logs, a eventual transferência internacional e a existência de avaliações de impacto. Os indicadores relevantes incluem a completude do mapeamento de dados, a existência e a atualização de RIPD, a identificação de grupos vulneráveis, a robustez dos controles de segurança, a ocorrência de incidentes, o grau de reidentificação residual e a suficiência das evidências apresentadas.

III. Dimensão de maturidade tecnológica

A dimensão de maturidade tecnológica, destina-se a acompanhar o nível de maturidade tecnológica da solução e de seus componentes críticos ao longo dos ciclos. Nesse eixo, devem ser utilizados indicadores como a variação do *Technology Readiness Level* (TRL) por ciclo, a maturidade dos elementos tecnológicos críticos, a aproximação entre o ambiente de teste e o ambiente operacional pretendido, as dependências técnicas

externas, a capacidade de integração com outros sistemas e a evolução da solução após devolutivas técnicas ou regulatórias.

IV. Dimensão de realismo experimental e benefícios esperados

A dimensão de realismo experimental e benefícios esperados busca comparar os usos esperados declarados pela empresa com os resultados observados nos testes, inclusive testes de extensão, testes de robustez, testes de cenários adversos e testes em condições distintas daquelas inicialmente previstas. Os indicadores devem medir a taxa de sucesso dos testes planejados, a manutenção ou alteração dos benefícios esperados, a ocorrência de resultados fora do escopo, a qualidade das justificativas apresentadas, o grau de generalização segura da solução e a eventual necessidade de restringir, redefinir ou condicionar o uso da aplicação.

V. Dimensão de riscos, mitigação e impacto residual

A dimensão de riscos, mitigação e impacto residual deve partir da identificação dos riscos iniciais, da classificação contínua de riscos, da matriz de probabilidade e impacto, do mapa de calor e do acompanhamento das medidas efetivamente implementadas. Os indicadores mínimos incluem número de riscos identificados por ciclo, distribuição dos riscos por severidade, variação do risco residual após mitigação, percentual de medidas concluídas, medidas planejadas ainda pendentes, riscos não mitigados, existência de gatilhos de descontinuidade e suficiência do plano de resposta a incidentes. Essa dimensão é central para verificar se a permanência da solução no *sandbox* reduz, estabiliza ou aumenta riscos a titulares de dados, usuários, parceiros, setores regulados e ao próprio ambiente regulatório.

VI. Dimensão de governança, transparência e aprendizagem regulatória

Na dimensão de governança, transparência e aprendizagem regulatória, os indicadores devem captar a existência de responsáveis identificados, mecanismos de *accountability*³, capacidade real de intervenção humana, documentação técnica e não técnica, explicabilidade das decisões, comunicação a usuários e stakeholders (partes interessadas), participação de autoridades setoriais e produção de conhecimento regulatório pela ANPD. Devem ser observados, entre outros, o número e a qualidade das lacunas regulatórias identificadas, a necessidade de coordenação com outros reguladores, a evolução dos critérios de avaliação, a utilidade das evidências produzidas para supervisão futura e a replicabilidade dos aprendizados para outros casos de IA.

Para garantir consistência metodológica, cada indicador deve ser registrado em ficha própria, contendo:

- nome do indicador;
- dimensão correspondente;
- definição;

³ Neste documento, *accountability* é compreendida como o princípio que reúne mecanismos de prestação de contas, transparência, demonstração de conformidade e responsabilização pelos tratamentos de dados pessoais.

- variável de origem, fórmula ou critério de aferição;
- escala de mensuração;
- fonte da evidência;
- periodicidade de atualização;
- responsável pela coleta;
- valor de linha de base;
- valor por ciclo, meta ou parâmetro de referência;
- interpretação do resultado; e
- limitações.

Quando o indicador for quantitativo, recomenda-se o uso de fórmulas simples, como percentual de atendimento de critérios, variação de pontuação entre ciclos, taxa de implementação de mitigação e produto entre probabilidade e impacto. Caso o indicador seja qualitativo, recomenda-se a adoção de categorias auditáveis, como “não iniciado”, “em desenvolvimento”, “implementado parcialmente”, “implementado com evidência suficiente” e “implementado com evidência insuficiente”.

A avaliação de impacto deve diferenciar três níveis de resultado. O primeiro nível corresponde aos produtos imediatos do ciclo, como fichas técnicas preenchidas, frameworks aplicados, pareceres emitidos, comentários recebidos, evidências anexadas e planos atualizados. O segundo nível corresponde aos resultados intermediários, como sugestões de melhoria da maturidade tecnológica, de redução de riscos, de aumento da transparência, de aprimoramento da governança e de adequação dos usos esperados. O terceiro nível corresponde ao impacto regulatório e institucional, entendido como a capacidade do *sandbox* de produzir critérios replicáveis, identificar lacunas normativas, orientar futuras decisões da ANPD, reduzir incertezas regulatórias e fortalecer a supervisão de sistemas de IA em contextos reais ou controlados. O principal objetivo é apresentar ao final do Ciclo 06 sugestões regulatórias consolidadas com base nos dados e informações coletadas nos testes.

Dessa forma, a definição de indicadores e variáveis deve permitir que o relatório não apenas descreva as atividades executadas, mas demonstre se a testagem produziu aprendizado regulatório, redução de riscos, aumento de maturidade, melhoria de conformidade e geração de evidências suficientes para fundamentar decisões sobre continuidade, restrição, adaptação, escalonamento ou descontinuidade da solução testada.

2.2 Desenho dos métodos de coleta e análise de dados referentes à testagem

A coleta e a análise dos dados no *Sandbox* Regulatório de IA e Proteção de Dados foram organizadas com base em um desenho metodológico progressivo, estruturado em ciclos sucessivos de testagem e em três fases operacionais: Fase A (Pré-Teste), Fase B (Testagem) e Fase C (Pós-Teste). A lógica adotada é cumulativa: a cada ciclo, as informações previamente coletadas são revisitadas, complementadas e, quando necessário, atualizadas em razão de alterações técnicas, regulatórias, organizacionais ou operacionais identificadas durante a execução dos testes. Assim, a metodologia não se limita à coleta inicial de documentos, mas acompanha a evolução da solução ao longo do tempo, registrando mudanças de premissas, ajustes de arquitetura, novas evidências, riscos emergentes, manifestações da ANPD, contribuições de stakeholders e resultados observados nas atividades práticas. Veja na Figura 3 as fases e as subfases.



Figura 3: Fases e subfases do ciclo de testagem.

Portanto, as fases estão detalhadas a seguir nos tópicos de 2.2.1 até 2.2.3, detalhando cada subfase envolvida no processo metodológico.

2.2.1 Fase A - Preparação e Detalhamento

A Fase A corresponde ao pré-teste e tem por finalidade estabelecer a base informacional mínima para que a solução possa ser analisada de maneira tecnicamente situada. Essa fase subdivide-se em A1, A2, A3 e A4.

A1. Detalhamento dos Participantes

A subfase A1 compreende o detalhamento da aplicação e dos participantes por meio da ficha técnica. Nessa etapa são coletadas informações sobre a arquitetura do modelo, versão, topologia, pipeline de inferência, inputs relevantes, uso padrão da ferramenta, cenário ideal de utilização, limitações de escopo, APIs utilizadas, fluxos de dados, controles de segurança, tratamento de dados pessoais e eventuais alterações

entre ciclos. Trata-se da principal fonte de coleta de dados para a compreensão técnico-funcional da solução e para a identificação de mudanças relevantes que possam afetar os testes subsequentes.

A2. Estágio de Desenvolvimento (TRL)

A subfase A2 trata do estágio de desenvolvimento do projeto, com base na matriz de maturidade tecnológica, especialmente a matriz TRL (Figura 4). Nessa etapa, são coletadas informações sobre o grau de maturidade da solução e de seus elementos tecnológicos críticos, permitindo acompanhar se a solução se mantém no estágio, evolui ou demanda reclassificação ao longo dos ciclos. A análise não se resume à declaração do nível de maturidade informado pela empresa; deve verificar a coerência entre o TRL declarado, as evidências apresentadas, o ambiente de demonstração, o ambiente relevante e o grau de estabilidade técnica observado durante os testes.



Figura 4: Escala TRL está organizada em 9 níveis.

A3. Coordenação Regulatória

No âmbito da subfase A3 refere-se à coordenação multiagente e à interoperabilidade regulatória. Seu objetivo é coletar dados sobre o arranjo institucional necessário à operação da solução, incluindo possíveis sobreposições de competência entre autoridades, interações com reguladores setoriais, riscos de conflito normativo, necessidade de instrumentos de cooperação e formas de harmonização entre exigências regulatórias. A análise dessa subfase verifica se a solução depende de coordenação com outros órgãos, se há riscos de sobreposição ou lacunas de competência e quais mecanismos podem ser utilizados para garantir alinhamento institucional.

A4. Medidas de Transparência

Durante a subfase A4 concentra-se em medidas de transparência e de conformidade regulatória. Nessa etapa são coletadas informações sobre explicabilidade, comunicação com usuários e titulares, documentação técnica e não técnica, requisitos do art. 20 da LGPD, eventuais tensões entre transparência, segurança técnica e segredo comercial, e as medidas adotadas para tornar a solução compreensível e contestável. A análise deve distinguir as camadas de explicação necessárias a diferentes públicos,

especialmente uma camada técnica, focada em auditores e reguladores, e outra funcional, dedicada a usuários e titulares.

2.2.2 Fase B - Experimentação e Gestão de Riscos

A Fase B corresponde à fase de teste propriamente dita e está subdividida em B1, B2 e B3.

B1. Aplicação de *Frameworks*

A subfase B1 trata do realismo experimental, dos usos esperados, dos benefícios e dos testes de extensão. Nessa etapa, são coletadas informações sobre a finalidade da solução, seus benefícios esperados, cenários de uso, testes planejados, testes executados e evidências produzidas. Conforme o tipo de solução, essa subfase pode envolver a aplicação de frameworks gerais, como o framework ético da UNESCO, frameworks específicos para o tipo de ferramenta, como referências aplicáveis a LLMs ou a sistemas de IA generativa, e frameworks setoriais, como os dedicados à saúde, à mobilidade ou ao setor financeiro. Também entram nessa subfase os testes operacionais definidos no cronograma, como testes com casos simulados ou reais controlados, testes de *red teaming*, testes de segurança, testes de pseudonimização, testes de geração de dados sintéticos, testes de integração, testes de execução algorítmica em ambiente controlado e avaliações qualitativas por especialistas.

É relevante destacar que foi selecionado o framework ético da UNESCO para aplicação em todas as empresas, devido à sua característica agnóstica e ao fato de o Brasil incorporar os princípios da UNESCO relacionados à IA em seus documentos internos. Adicionalmente, houve a participação do consultor internacional João Pita Costa, o qual integra o IRCAI UNESCO (Centro de IA da UNESCO, localizado na Eslovênia) e está atuando ativamente na adequação do framework às empresas participantes do *sandbox*.

B2. Classificação de Riscos

A subfase B2 corresponde à classificação contínua de riscos. Ela se desdobra em três pareceres principais: parecer sobre usos esperados, benefícios e testes de extensão; parecer sobre mapa de calor e mitigações; e parecer sobre mitigações efetivadas e planejadas. O primeiro parecer analisa se os usos inicialmente declarados permanecem adequados após a execução dos testes, se os benefícios esperados foram mantidos, restringidos ou reformulados, e se os testes de extensão revelaram novas limitações. O segundo parecer organiza os riscos identificados em um mapa de calor, permitindo visualizar a distribuição desses riscos por probabilidade, impacto e severidade. O terceiro parecer verifica o que foi efetivamente mitigado, o que está em andamento, o que permanece apenas planejado e o que não foi comprovado por evidências suficientes.

B3. Acompanhamento e *Feedbacks*

Na subfase B3 corresponde ao acompanhamento regulatório e participativo. Ela inclui a coleta dos comentários da ANPD na etapa de revisão e os ajustes subsequentes realizados pelas empresas ou pelas equipes técnicas. Conforme o cronograma, a revisão da ANPD é tratada como uma etapa própria, seguida de um prazo específico para ajustes. Essa subfase também pode incluir comentários de stakeholders, reuniões, workshops com empresas, interação com inscritos em novos editais e diálogos com autoridades setoriais. Nesse sentido, o objetivo é permitir que a coleta de dados incorpore não apenas documentos formais e resultados técnicos, mas também percepções regulatórias, dúvidas interpretativas, preocupações institucionais e contribuições externas relevantes.

2.2.3 Fase C - Avaliação e Ajuste

A Fase C corresponde ao pós-teste e está subdividida em C1, C2 e C3.

C1. Revisão de critérios

A subfase C1 envolve a revisão de critérios e a elaboração de pareceres sobre os êxitos a serem relatados pela ANPD e pelas empresas participantes. Trata-se de uma oportunidade para registrar os ganhos e avanços constatados em cada Ciclo de Testagem, permitindo incorporar ajustes para o próximo ciclo a iniciar.

O parecer sobre êxitos a ser preenchido pela Comissão de *Sandbox* da ANPD registra as lições regulatórias extraídas do ciclo, incluindo lacunas normativas, critérios replicáveis que se mostraram úteis, evidências dos riscos detectados, limitações técnicas observadas, pontos que exigem coordenação interinstitucional e parâmetros que podem orientar futuras ações de supervisão.

Por outro lado, o parecer sobre êxitos, a ser preenchido pelas empresas, visa a coletar registros de aprendizados operacionais, melhorias de conformidade regulatória, dúvidas resolvidas ao longo da testagem, riscos mitigados, identificar eventuais obstáculos regulatórios e/ou operacionais à inovação e condições para a continuidade ou o aprimoramento da solução. As empresas não serão obrigadas a responder a todos os pontos, mas essas informações são relevantes, inclusive para futuros *sandboxes*, pois se trata de material que evidencia incentivos para que novas empresas participem de potenciais chamadas. Além disso, permite que as autoridades reguladoras compreendam possíveis ganhos que, sob a ótica das empresas, são relevantes e podem se estender a eventuais novos participantes.

C2. Plano de Descontinuidade

A subfase C2 trata do Plano de Descontinuidade e, quando aplicável, de atividades adicionais previstas no ciclo final, como auditoria algorítmica. O Plano de Descontinuidade é coletado e atualizado como documento vivo, com registro de hipóteses de encerramento, suspensão, reversão, migração, eliminação, anonimização, retenção controlada, comunicação aos usuários e tratamento de dados remanescentes. A análise

dessa subfase verifica se a solução possui condições seguras de encerramento ou de transição, se os riscos residuais foram tratados e se há evidências suficientes para demonstrar que eventual descontinuidade não produzirá impactos indevidos sobre titulares, usuários, parceiros, sistemas internos ou bases de dados.

C3. Relatório do Ciclo

Na subfase C3 corresponde ao relatório do ciclo. Nessa etapa, os dados coletados nas fases A, B e C são consolidados em relatório parcial, contemplando as atividades executadas, os documentos recebidos, os testes realizados, as evidências analisadas, os riscos identificados, as mitigações observadas, os comentários da ANPD, os pontos positivos, os pontos negativos, as pendências e as recomendações. Ao final dos ciclos, os relatórios parciais são consolidados em um relatório final, acompanhado da documentação dos resultados dos testes. Essa consolidação permite comparar o percurso completo de cada solução, desde sua caracterização inicial até os resultados e os aprendizados regulatórios gerados.

O método de análise dos dados adota uma estratégia de triangulação. As informações declaradas pelas empresas são comparadas às evidências documentais apresentadas, aos registros técnicos, aos resultados dos testes, às respostas aos formulários, aos pareceres das equipes especializadas, aos comentários da ANPD e, quando houver, às contribuições de stakeholders. Essa triangulação permite reduzir a dependência da autodeclaração empresarial e qualificar a análise com base em evidências verificáveis. Sempre que houver divergência entre declaração, evidência e resultado observado, a inconsistência deve ser registrada como ponto de atenção, pendência ou condicionante para o ciclo seguinte.

As seis dimensões da matriz de avaliação são utilizadas apenas como estrutura de organização analítica dos dados coletados nas fases. A Fase A alimenta principalmente as dimensões técnico-funcional, de dados, de privacidade e segurança, de maturidade tecnológica, de governança e de transparência. A Fase B concentra-se principalmente nas dimensões de realismo experimental, benefícios esperados, riscos, mitigação e impacto residual. A Fase C organiza os dados relativos à aprendizagem regulatória, às condições de continuidade ou de descontinuidade, à revisão de critérios e à consolidação dos resultados. Desse modo, as dimensões funcionam como lentes de classificação, permitindo que os dados coletados em cada fase sejam analisados de forma coerente e comparável.

A análise combina leitura qualitativa, verificação técnica e classificação estruturada. A leitura qualitativa é aplicada às justificativas, pareceres, comentários, relatos de reuniões, avaliações de especialistas e explicações fornecidas pelas empresas. A verificação técnica incide sobre artefatos como fluxos de dados, logs, registros de API, documentação de segurança, resultados de testes, matrizes TRL, mapas de risco e evidências de mitigação. A classificação estruturada é utilizada para organizar o status das medidas, distinguindo controles efetivados, em andamento, planejados ou não comprovados, bem como para registrar a evolução das pendências entre ciclos.

Com essa metodologia, cada ciclo de testagem produz três tipos de informação: primeiro, dados descritivos sobre a solução e seu contexto; segundo dados observacionais

sobre o comportamento da solução durante os testes; e terceiro, dados avaliativos sobre riscos, mitigação, adequação, transparência, governança e aprendizagem regulatória. O resultado esperado não é apenas verificar se a solução funciona tecnicamente, mas compreender como ela se comporta em ambiente controlado, quais riscos emergem ou são reduzidos, quais evidências sustentam sua continuidade e quais aprendizados podem orientar futuras decisões regulatórias da ANPD, garantindo o equilíbrio entre a inovação e a proteção dos direitos fundamentais. Por fim, torna-se relevante construir um ambiente de confiança mútua entre regulador e regulado, o qual é trabalhado nos encontros, workshops e eventos organizados.

2.3 Consolidação dos dados e estruturação do plano de monitoramento e avaliação

A consolidação dos dados e a estruturação do plano de monitoramento e avaliação serão realizadas com base na organização progressiva das informações coletadas ao longo dos ciclos de testagem do *Sandbox* Regulatório de IA e Proteção de Dados (6 ciclos). Essa etapa tem por finalidade transformar os documentos, evidências, pareceres, comentários e resultados produzidos em cada fase de teste em uma base analítica coerente, capaz de apoiar tanto o acompanhamento das soluções participantes quanto a formulação de recomendações para a ANPD.

A consolidação partirá das informações obtidas nas fases A, B e C do cronograma de testagem.

I. Fase A

Na Fase A, serão sistematizados os dados de caracterização inicial e de atualização técnica das soluções, incluindo a ficha técnica da aplicação, a arquitetura do modelo, o pipeline de inferência, as APIs utilizadas, os fluxos de dados, as limitações de escopo, a matriz de maturidade tecnológica e eventuais alterações ocorridas entre ciclos. Essas informações permitirão estabelecer a linha de base de cada solução e identificar, desde o início, os aspectos técnicos, funcionais, regulatórios e organizacionais a serem monitorados nos ciclos subsequentes.

II. Fase B

Na Fase B, a consolidação incorporará os dados provenientes da execução dos testes propriamente ditos, dos testes de extensão, da análise dos usos esperados e dos benefícios declarados, da aplicação de frameworks de avaliação e dos pareceres técnicos sobre riscos e mitigações. Nessa fase, serão organizados os resultados observados durante os testes, as divergências entre o comportamento esperado e o verificado, os riscos identificados, os mapas de calor, as medidas de mitigação efetivadas, as medidas planejadas, as pendências e as condicionantes eventualmente necessários à continuidade da testagem. Também serão considerados os comentários da ANPD, as

devolutivas da equipe técnica e, quando aplicável, as contribuições de stakeholders ou de autoridades setoriais.

III. Fase C

Na Fase C, a consolidação terá natureza avaliativa e conclusiva. Serão reunidos os pareceres de encerramento do ciclo, os relatórios de êxito para a ANPD e para as empresas, o Plano de Descontinuidade, os registros de aprendizagem regulatória e as recomendações sobre continuidade, restrição, adaptação da solução testada. O objetivo será verificar não apenas se os testes foram executados, mas também se produziram evidências suficientes para apoiar decisões regulatórias, aperfeiçoar a governança das soluções e gerar parâmetros replicáveis para futuras iniciativas de supervisão de sistemas de IA.

O plano de monitoramento e avaliação será estruturado como um instrumento cumulativo. A cada ciclo, as informações recebidas serão classificadas, analisadas e comparadas com os dados anteriormente consolidados. Essa comparação permitirá identificar alterações relevantes na arquitetura da solução, no tratamento de dados pessoais, nos controles de segurança, na maturidade tecnológica, nos riscos residuais, nas medidas de transparência, na governança interna e na relação com autoridades regulatórias. Quando houver mudanças relevantes de finalidade, fornecedor, modelo, base de dados, fluxo de integração, interface, nível de risco ou forma de interação com usuários e titulares, essas alterações deverão ser registradas como pontos de atenção para análise no ciclo seguinte.

Dessa forma, recorda-se que a estruturação do plano deverá observar seis dimensões analíticas apenas como forma de organizar os dados coletados: técnico-funcional; dados, privacidade e segurança; maturidade tecnológica; realismo experimental e benefícios esperados; riscos, mitigação e impacto residual; e governança, transparência e aprendizagem regulatória. Essas dimensões servirão para distribuir os elementos das fases de teste, evitando que a consolidação se limite a uma sequência documental. Assim, cada documento recebido, parecer emitido, teste executado ou comentário formulado será associado à dimensão correspondente, facilitando a leitura integrada dos resultados.

O processo de análise combinará três movimentos metodológicos:

- O primeiro será a verificação documental, destinada a confirmar se as informações exigidas foram apresentadas, se as evidências foram anexadas e se os registros estão completos, atualizados e compatíveis com a fase do ciclo.
- O segundo será a análise técnica e regulatória, voltada a examinar a suficiência das evidências, a coerência entre declarações e testes executados, a adequação das medidas de mitigação e a necessidade de condicionantes.

- O terceiro será a consolidação avaliativa, na qual os achados de cada ciclo serão sintetizados em pareceres, comentários e relatórios parciais, a fim de alimentar o relatório final do *sandbox*.

Os pareceres técnicos terão papel central nesse processo. Eles deverão registrar os principais achados de cada ciclo, bem como distinguir pontos positivos, pontos negativos, lacunas, riscos, medidas adotadas, medidas pendentes e recomendações. Sempre que pertinente, os pareceres deverão indicar se determinada medida foi efetivamente implementada, se está em andamento, se permanece apenas planejada ou se não foi comprovada por evidência suficiente. Essa classificação permitirá que o plano de monitoramento acompanhe a evolução das soluções de maneira objetiva, sem depender exclusivamente de autodeclarações das empresas participantes.

Os comentários da ANPD e das equipes técnicas serão incorporados como insumos para revisão e reorientação do processo de testagem. Quando os comentários indicarem necessidade de complementação documental, ajuste de escopo, reforço de segurança, esclarecimento regulatório, revisão de transparência, aprimoramento de governança ou nova rodada de testes, tais demandas deverão ser registradas como pendências monitoráveis no ciclo seguinte. A consolidação dos comentários permitirá preservar a rastreabilidade entre a observação formulada, a resposta apresentada pela empresa, a evidência produzida e a conclusão adotada no parecer subsequente.

Os relatórios de êxito terão função complementar aos pareceres técnicos. Enquanto os pareceres examinam a adequação da solução, os riscos e as condicionantes, os relatórios de êxito deverão registrar os aprendizados decorrentes da testagem. Para a ANPD, esses relatórios deverão sistematizar lacunas regulatórias, critérios de supervisão, parâmetros mínimos de documentação, padrões de risco observados, desafios de coordenação institucional e hipóteses de aprimoramento regulatório. Para as empresas, deverão indicar os avanços obtidos, as melhorias de governança, os aperfeiçoamentos técnicos, os controles implementados, as evidências fortalecidas e as condições para a continuidade segura da solução.

Adicionalmente, ao encerramento da Fase B de cada ciclo, será convocada uma reunião de devolutiva com os participantes das empresas e com a Comissão de *Sandbox* da ANPD, com a finalidade de apresentar e discutir os dados coletados, os documentos analisados, os testes realizados e os resultados dos pareceres técnicos relativos ao respectivo ciclo. Essa reunião terá a função de validação e alinhamento metodológicos, permitindo esclarecer eventuais lacunas, registrar comentários da ANPD e das empresas, indicar pendências para o ciclo subsequente e assegurar a rastreabilidade entre as evidências apresentadas, os achados técnicos e as recomendações formuladas.

No caso das testagens realizadas com o apoio da ferramenta Harpia⁴ da USP, será necessário preservar, no início de cada rodada de testes, a versão da solução submetida como referência do ciclo. Assim, não deverão ser realizadas atualizações de modelo, interface, arquitetura, integrações, prompts, fluxos de dados ou quaisquer outros

⁴ Mais informações disponíveis em: <https://sites.usp.br/keml/en/harpia-2/>.

componentes capazes de alterar o comportamento da solução durante a execução dos testes, sob pena de comprometer a comparabilidade e a confiabilidade dos resultados aferidos. Eventuais alterações indispensáveis deverão ser previamente comunicadas, justificadas e registradas, podendo ensejar a reexecução, a segregação ou a reclassificação dos testes afetados.

Por fim, a consolidação dos dados deverá gerar um relatório final integrado, elaborado a partir dos relatórios parciais de cada ciclo. Esse relatório final deverá apresentar a trajetória das soluções testadas, os principais resultados obtidos, os riscos identificados e tratados, as pendências remanescentes, os êxitos alcançados, as recomendações à ANPD e os aprendizados regulatórios extraídos da experiência. Com isso, o plano de monitoramento e avaliação permitirá que o *sandbox* seja compreendido não apenas como um ambiente de teste de soluções específicas, mas também como um mecanismo estruturado de produção de evidências, de aprendizagem institucional e de fortalecimento da governança regulatória em inteligência artificial e proteção de dados.

3 Plano de *Sandbox* e Plano de Descontinuidade

Durante a fase de nivelamento, as empresas participantes com o apoio do CIAAM e revisão da Comissão de *Sandbox* elaboraram o Plano de *Sandbox* (PdS) e o Plano de Descontinuidade. Esses dois documentos desempenham papel estruturante para a fase de testagem.

3.1 Plano de *Sandbox* (PdS)

O PdS é caracterizado como o roteiro que guiará todo o experimento do projeto-piloto, estabelecendo o acordo de execução entre a empresa participante e a ANPD. O objetivo principal do PdS é permitir que a inovação da empresa seja representada em etapas concretas, com a possibilidade de serem mensuráveis e auditáveis para testes.

O PdS é construído em cima dos pilares de Segurança Jurídica, Validação de Conformidade, Gestão de Expectativas e Recursos e Preparação para o Mercado, que estão apresentados na Figura 5.

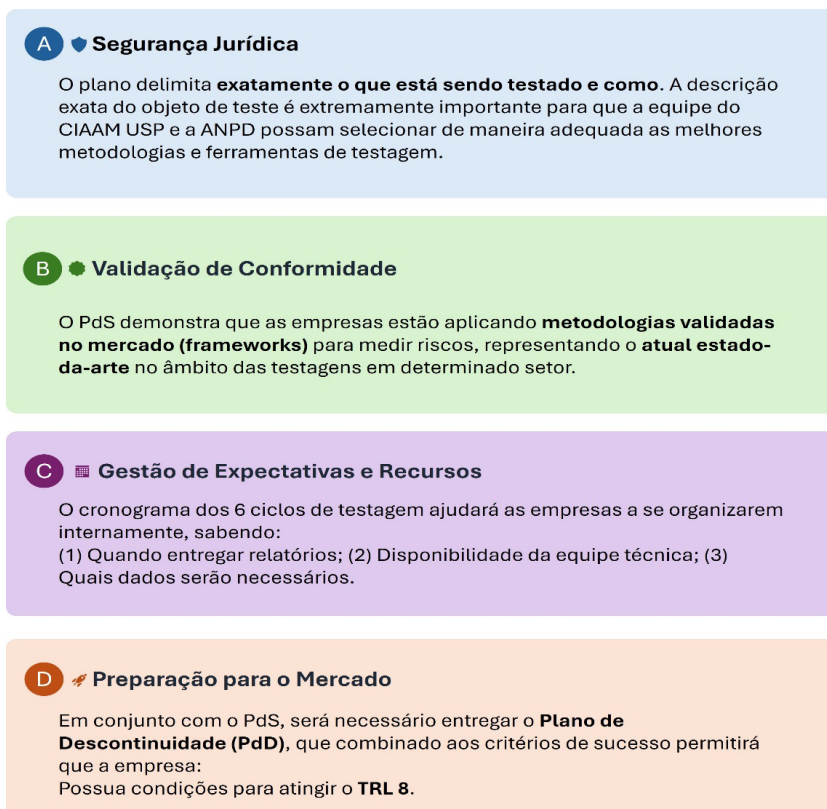


Figura 5: Quatro pilares do Plano de *Sandbox*.



Figura 6: Seções do Plano de *Sandbox*.

Para que o PdS atenda ao seu objetivo, ele deve ser elaborado contendo informações claras e objetivas que servirão como a linha de base para o início dos ciclos de teste. É importante destacar que as informações declaradas são vinculantes à execução do projeto após a aprovação pela Comissão de *Sandbox* da ANPD.

Dessa forma, o PdS é dividido em 7 seções, que contemplam:

- Plano Operacional e Perfil Técnico: estabelece as diretrizes para a execução do projeto-piloto, descrevendo de forma estruturada a arquitetura da solução, os procedimentos de teste, o fluxo de dados e os mecanismos de processamento utilizados pelo sistema. Além disso, contempla a declaração do nível de maturidade tecnológica (TRL) da solução, permitindo avaliar sua prontidão no início de cada ciclo de testagem, bem como o cronograma operacional dos seis ciclos previstos, com a definição de objetivos, etapas, marcos e ambiente controlado de experimentação. Em conjunto, esses elementos asseguram a padronização da execução dos testes, a rastreabilidade das atividades e a avaliação consistente da evolução técnica da solução ao longo do *sandbox* regulatório;
- Análise de Riscos e Impactos: Apresenta os principais riscos associados à solução, sua classificação quanto à probabilidade e ao impacto, bem como as medidas de mitigação adotadas para reduzir seus efeitos ao longo dos ciclos de testagem. Também devem ser descritos os protocolos de segurança implementados e o mapeamento das normas e autoridades reguladoras aplicáveis ao projeto, evidenciando a conformidade regulatória, a proteção dos dados pessoais e o acompanhamento contínuo da evolução dos riscos durante a execução do projeto;
- Conformidade e Governança: Estabelece os critérios objetivos para acompanhamento da evolução do projeto, definindo métricas de sucesso, indicadores técnicos, critérios de conformidade e requisitos

para elaboração dos relatórios de cada ciclo de testagem. Esses elementos serão revisados junto à ANPD no Ciclo 1 e orientarão a avaliação da efetividade, segurança e viabilidade da solução ao longo do projeto;

- **Proteção de Dados:** estabelece os critérios para a proteção de segredos comerciais e industriais no contexto de transparência controlada do Sandbox. A empresa deverá identificar de forma precisa as informações confidenciais, justificar o potencial dano decorrente de sua divulgação e diferenciar conteúdos passíveis de publicação daqueles que devem permanecer restritos. Dessa forma, busca-se equilibrar a transparência dos resultados do projeto com a preservação da inovação, garantindo que a ANPD possa divulgar aprendizados relevantes sem comprometer vantagens competitivas legítimas;
- **Medidas de Transparência:** As Medidas de Transparência devem demonstrar como a empresa garantirá informações claras sobre a participação da solução no projeto, os critérios utilizados em decisões automatizadas e os mecanismos de explicabilidade adotados. Além disso, devem ser apresentados os frameworks éticos, técnicos e setoriais que orientarão a validação da solução, evidenciando o compromisso com a transparência, a conformidade regulatória e a proteção dos direitos dos titulares de dados;
- **Plano de Descontinuidade:** O Plano de Descontinuidade estabelece as estratégias e procedimentos para o encerramento seguro do projeto, contemplando cenários planejados ou emergenciais. O Plano de Descontinuidade é detalhado na Seção 3.2; e Anexos e Documentação Complementar: Os anexos complementam o plano por meio de evidências técnicas que apoiam a avaliação da solução proposta. Devem contemplar o diagrama da arquitetura e do fluxo de dados, a matriz de risco e o diagrama de Gantt dos seis ciclos de testagem, permitindo avaliar aspectos de segurança, gestão de riscos e planejamento do projeto. Também poderão ser incluídos anexos facultativos que sejam pertinentes e agreguem valor ao processo de avaliação.

Mais informações sobre o Plano de *Sandbox* estão disponíveis em <https://sites.usp.br/sandbox/plano-de-sandbox/>.

3.2 Plano de Descontinuidade (PdD)

O PdD é um instrumento fundamental para garantir a segurança, transparência e inovação responsável de sistemas de inteligência artificial em ambientes de teste. O objetivo do Plano de Descontinuidade é estabelecer as diretrizes e os procedimentos para

o encerramento das atividades, com especial atenção ao tratamento de dados pessoais, bem como à mitigação dos eventuais riscos decorrentes da descontinuação ou do encerramento do projeto.

De forma detalhada, o primeiro objetivo do Plano de Descontinuidade é definir o papel desse instrumento no contexto do *sandbox* regulatório. Nesse sentido, o plano estabelece as estratégias necessárias para que o encerramento das atividades ocorra de forma segura, planejada e responsável, descrevendo as etapas que devem ser executadas durante o processo de descontinuação. Ao estruturar esse fluxo de maneira clara, o plano busca assegurar a proteção dos dados pessoais, promover a transparência das ações adotadas e preservar os princípios da inovação responsável ao término do projeto.

O segundo objetivo consiste em garantir o atendimento aos requisitos mínimos estabelecidos no Edital do *Sandbox* Regulatório da ANPD, especialmente aqueles previstos no art. 24, § 2º. Para tanto, o Plano de Descontinuidade deve contemplar, no mínimo, o cronograma de encerramento das atividades, o plano de tratamento dos dados pessoais após a conclusão do projeto, as medidas destinadas à proteção dos direitos dos titulares, os mecanismos de auditoria e validação da conformidade, bem como a identificação e mitigação dos riscos que possam comprometer a execução do encerramento.

O terceiro objetivo é fornecer orientações práticas que facilitem a elaboração e a implementação do Plano de Descontinuidade pelas instituições participantes. Para isso, são disponibilizados protocolos, listas de verificação, modelos e outros instrumentos de apoio que auxiliam na execução das atividades previstas. Esses recursos contribuem para padronizar procedimentos, reduzir inconsistências e assegurar que o encerramento do projeto seja conduzido de forma consistente, transparente e em conformidade com as obrigações regulatórias estabelecidas para o ambiente experimental.

Mais informações sobre o Plano de Descontinuidade estão disponíveis em <https://sites.usp.br/sandbox/plano-de-descontinuidade/>.

