HIGOR VINICIUS NOGUEIRA JORGE

IA íntegra

Segurança, Ética e Responsabilidade no Uso da Inteligência Artificial

2026

EDITORA JusPODIVM
www.editorajuspodivm.com.br

CAPÍTULO 7

COMO CONSTRUIR IA SEGURA: BOAS PRÁTICAS DE DESENVOLVIMENTO

Há uma crença persistente de que segurança e conformidade são problemas que se resolvem depois – que o desenvolvimento de um sistema de IA é tarefa dos engenheiros, e que as perguntas sobre responsabilidade, viés e proteção de direitos entram em cena apenas quando algo dá errado. Essa crença é, simultaneamente, equivocada e cara.

Cara porque corrigir um sistema viciado depois de implantado – retreiná-lo, reauditar seus resultados, reparar os danos que causou, reconstruir a confiança institucional que perdeu – custa consistentemente mais do que teria custado projetar corretamente desde o início. E equivocada porque as decisões mais consequentes sobre segurança, viés e prestação de contas não são tomadas no momento do incidente; são tomadas no momento em que alguém escolhe os dados de treino, define o objetivo do modelo, decide o que o sistema deve otimizar e determina quais salvaguardas serão incorporadas à arquitetura.

Segurança em IA começa no projeto, não na resposta ao problema. Esse é o princípio central deste capítulo – e é de onde derivam todas as práticas que o seguem.

Segurança desde o projeto

O conceito de segurança desde o projeto tem raízes na engenharia de software e na proteção de dados, mas sua aplicação à IA tem especificidades que merecem atenção.

Em software tradicional, projetar segurança desde o início significa pensar, antes de escrever código, em como o sistema pode ser atacado, como dados podem ser roubados, como o acesso pode ser comprometido. As perguntas são estruturais: quem pode acessar o quê? Como o sistema se comporta quando recebe entrada inesperada? O que acontece se um componente falha?

Em IA, essas perguntas continuam válidas – e são acrescidas de outras, específicas da natureza dos sistemas inteligentes. Com que dados o modelo será treinado, e esses dados são representativos da população que o sistema vai afetar? Que variáveis serão incluídas, e alguma delas é proxy de característica protegida? Qual é o objetivo que o modelo vai otimizar, e esse objetivo captura adequadamente o que se quer alcançar sem criar incentivos perversos? O modelo precisa explicar suas decisões, e a arquitetura escolhida permite isso?

Essas perguntas não têm respostas técnicas automáticas. São perguntas de julgamento – e o julgamento que elas exigem combina competência técnica com compreensão dos princípios éticos e jurídicos discutidos nos capítulos anteriores. O engenheiro que nunca se perguntou sobre viés não vai detectá-lo nos dados. O gestor que nunca entendeu o que é uma caixa-preta não vai exigir explicabilidade como requisito. A integração entre perspectiva técnica e perspectiva jurídico-ética não é luxo – é condição para que o produto final seja responsável.

Modelagem de ameaças antes do primeiro código

Antes de qualquer linha de código, equipes que desenvolvem IA responsável realizam um exercício sistemático de antecipação: quem tem interesse em comprometer este sistema? De que formas poderia fazê-lo? Quais seriam as consequências para os usuários, para os afetados pelas decisões do sistema, para a organização?

Esse exercício – chamado de modelagem de ameaças – é familiar na segurança da informação tradicional e precisa ser adaptado para IA. Além das ameaças convencionais de invasão e roubo de dados, a modelagem de ameaças para IA deve contemplar: a possibilidade de envenenamento do conjunto de treino; a possibilidade de que o sistema seja usado de forma diferente da prevista, em contextos para os quais não foi validado; a possibilidade de que grupos específicos sejam sistematicamente prejudicados por seu funcionamento; e a possibilidade de que o sistema seja manipulado por entradas adversarialmente construídas.

A modelagem de ameaças não precisa ser exercício formal e burocrático para ser útil. Em sua forma mais simples, é uma conversa estruturada entre desenvolvedores, gestores e, idealmente, representantes dos grupos que serão afetados pelo sistema: o que pode dar errado? Quem sai prejudicado se der? Como detectaríamos que deu? O que faríamos?

O resultado dessa conversa deve ser documentado – e deve orientar as decisões de design que vêm a seguir.

Pipeline de dados: onde tudo começa

O conjunto de dados de treino é o fundamento sobre o qual o modelo aprende. Problemas introduzidos aqui se propagam para tudo que vem depois – e frequentemente são invisíveis no resultado final, porque o modelo aprendeu a errar de forma consistente, sem sinais óbvios de falha.

Um pipeline de dados responsável começa pela verificação de proveniência: de onde cada dado veio? Com qual autorização foi coletado? Para qual finalidade original? Essa verificação não é apenas exigência de proteção de dados – é condição para que a cadeia de custódia do modelo seja defensável. Em contexto de investigação criminal, a pergunta sobre proveniência é ainda mais crítica: dados coletados sem amparo legal contaminam o modelo e, por extensão, qualquer decisão que ele embasar.

A seguir, verificação de representatividade: o conjunto de dados reflete adequadamente a diversidade da população que o sistema vai afetar? Grupos sub-representados no treino serão sistematicamente mal servidos pelo modelo. Essa verificação exige não apenas olhar para o tamanho do conjunto, mas para sua composição – quantos exemplos de cada grupo demográfico relevante? Com que qualidade e variação?

O armazenamento dos dados de treino merece a mesma proteção que qualquer dado pessoal sensível: criptografia em repouso e em trânsito, controle de acesso rigoroso com logs auditáveis, backups seguros, e retenção apenas pelo tempo necessário. Dados de treino são frequentemente tratados como ativo técnico interno – menos protegidos do que dados de produção – quando na verdade contêm informações sensíveis de pessoas reais, com todas as implicações jurídicas que isso implica.

A separação entre dados de treino, dados de validação e dados de teste é requisito técnico básico, mas merece menção: usar os mesmos dados para treinar e avaliar o modelo produz métricas de desempenho artificialmente otimistas que não se traduzem em desempenho real. Esse erro – tecnicamente chamado de vazamento de dados – é mais comum do que deveria ser, e suas consequências aparecem apenas em produção, quando o modelo encontra dados que genuinamente nunca viu.

Treinamento responsável

O processo de treinamento de um modelo de IA precisa ser documentado de forma que seja possível, em qualquer momento futuro, responder às perguntas que a responsabilidade exige: com que dados esse modelo foi treinado? Quando? Por quem? Quais eram os vieses conhecidos dos dados? Que decisões foram tomadas sobre arquitetura, objetivos e parâmetros, e por quê?

Essa documentação tem nome estabelecido na literatura técnica: cartão do modelo – um passaporte do modelo que registra seu propósito, suas condições de uso adequado, suas limitações conhecidas, suas métricas de desempenho desagregadas por grupo demográfico, e as condições sob as quais não deve ser usado. A analogia com a bula de medicamento não é forçada: assim como um remédio eficaz pode ser perigoso quando usado para indicação errada ou por pessoa com contraindicação específica, um modelo de IA preciso em média pode ser inadequado para determinados subgrupos ou contextos.

A documentação do treinamento também inclui versionamento: cada versão do modelo – com os dados usados, os parâmetros escolhidos e as métricas obtidas – deve ser registrada e preservada. Se em algum momento futuro for necessário determinar qual versão do modelo tomou qual decisão, em qual data, essa informação precisa estar disponível. Sem versionamento rigoroso, a rastreabilidade que o princípio da prestação de contas exige é impossível.

Testes que não podem ser pulados

Antes de qualquer implantação que afete decisões sobre pessoas reais, um conjunto mínimo de testes precisa ser realizado e documentado.

Testes de viés verificam se o modelo produz resultados sistematicamente diferentes para grupos demográficos distintos. Métricas de equidade – como paridade de precisão, igualdade de taxas de falso positivo e falso negativo entre grupos, e análise de impacto desproporcional – devem ser calculadas e avaliadas antes da implantação, com limites predefinidos para o que é considerado aceitável. Se o modelo não passa nesses testes, não está pronto para operar em contexto que afeta direitos.

Testes de robustez verificam como o sistema se comporta diante de entradas fora do esperado: dados ruidosos, entradas incompletas, casos extremos que nunca apareceram no treino. Um sistema que funciona perfeitamente em condições ideais e falha catastroficamente em condições reais – que são sempre menos ideais – não está pronto para produção. Testes de robustez também incluem, para sistemas de maior risco, testes de resistência a ataques adversariais: o sistema mantém seu comportamento quando recebe entradas cuidadosamente construídas para enganá-lo?

Testes de explicabilidade verificam se o sistema é capaz de produzir, para cada decisão relevante, uma explicação que seja compreensível para o profissional que a usará e contestável por quem for afetado. Isso pode significar verificar se um modelo interpretável produz regras de decisão legíveis, ou se as ferramentas de explicação pós-hoc aplicadas a um modelo mais complexo produzem resultados estáveis e coerentes.

Testes de performance desagregada verificam se a acurácia global esconde disparidades por grupo – o ponto discutido no capítulo anterior sobre a ilusão da acurácia. Um sistema com 95% de acurácia global pode ter 70% de acurácia para um grupo específico, o que é inaceitável em qualquer decisão que afete direitos daquele grupo.

O registro dos testes realizados, seus resultados e as decisões tomadas com base neles é parte da cadeia de documentação que

sustenta a diligência prévia. Sem ele, não há como demonstrar que a implantação foi precedida de avaliação responsável.

Explicabilidade como requisito

A tensão entre acurácia e explicabilidade – modelos mais precisos tendem a ser menos interpretáveis – é real, mas frequentemente superestimada como justificativa para o uso irrestrito de caixas-pretas. Em muitos contextos, a diferença de acurácia entre um modelo interpretável adequadamente projetado e um modelo de aprendizado profundo opaco é menor do que os defensores das caixas-pretas sugerem – e o custo em termos de contestabilidade e responsabilidade é enorme.

A recomendação prática, especialmente para sistemas que embasam decisões que afetam liberdade, saúde ou crédito, é adotar o modelo mais interpretável que atenda aos requisitos de desempenho. Se um modelo de árvore de decisão ou de regressão logística produz acurácia suficiente para o caso de uso, não há justificativa técnica – apenas de conveniência – para adotar rede neural profunda inexplicável no lugar.

Quando a complexidade do problema genuinamente exige modelos mais sofisticados, existem abordagens que preservam algum grau de explicabilidade: arquiteturas com atenção que indicam quais partes do input mais influenciaram o output; técnicas de atribuição de importância de features que mostram quais variáveis contribuíram para a decisão; e ferramentas de explicação local que aproximam o comportamento do modelo por explicações compreensíveis em casos específicos. Essas ferramentas têm limitações importantes – suas explicações são aproximações, não raciocínios reais do modelo – mas são melhores do que a ausência completa de explicação quando a arquitetura do modelo não permite outro caminho.

O critério operacional para escolha de arquitetura deve incluir, de forma explícita, a pergunta: se este sistema produzir uma recomendação que resulte em dano a uma pessoa, consigo explicar a um magistrado por que o sistema recomendou o que recomendou? Se a resposta for não, a arquitetura precisa ser reconsiderada ou a supervisão humana precisa ser estruturalmente reforçada.

Versionamento e rastreabilidade

Um sistema de IA em produção não é um objeto estático. É atualizado, retreinado com dados novos, ajustado em seus parâmetros, modificado para corrigir erros. Cada uma dessas mudanças pode alterar seu comportamento de formas que não são imediatamente óbvias – e que só serão relevantes quando alguém precisar determinar qual versão do sistema tomou qual decisão, em qual data, com qual comportamento esperado.

Versionamento rigoroso responde a essa necessidade. Código, dados de treino, parâmetros do modelo e configurações de implantação devem ser versionados de forma que seja possível reproduzir exatamente qualquer versão anterior do sistema. Logs de decisão – registros de cada input que o sistema recebeu, o output que produziu, a versão do modelo que estava em uso e o operador humano que revisou ou executou a decisão – devem ser preservados pelo tempo adequado ao tipo de decisão que o sistema embasou. Em contexto de investigação criminal, o prazo de preservação deve ser compatível com os prazos processuais aplicáveis.

Essa infraestrutura de rastreabilidade pode parecer onerosa do ponto de vista operacional. É. Mas ela é a condição para que a prestação de contas seja possível – e, na ausência de um incidente, é justamente o que permite demonstrar que o sistema operou dentro dos parâmetros documentados. É, em outros termos, tanto o instrumento de responsabilização quando as coisas dão errado quanto a evidência de responsabilidade quando as coisas funcionam bem.

Monitoramento em produção

A implantação não é o ponto final do processo responsável de desenvolvimento – é o início de uma fase diferente, mas igualmente exigente. Sistemas de IA em produção precisam ser monitorados de forma contínua e sistemática.

Monitoramento de desempenho verifica se as métricas que o sistema apresentava antes da implantação se mantêm em produção. Derrapagem de modelo – discutida no capítulo anterior – pode ocorrer de forma gradual e imperceptível; apenas comparação sistemática entre predições do modelo e resultados reais ao longo do tempo permite detectá-la antes que se torne grave.

Monitoramento de equidade verifica se as disparidades de desempenho entre grupos, medidas antes da implantação, se mantêm estáveis ou se novas disparidades emergem com dados de produção. O viés não é estático: pode se amplificar com o uso, especialmente em sistemas que aprendem continuamente.

Monitoramento de uso verifica se o sistema está sendo usado dentro dos parâmetros para os quais foi projetado e validado. Um sistema desenvolvido para determinado contexto e usado em outro pode produzir resultados completamente inadequados – não por falha do modelo, mas por inadequação de escopo. Esse é um dos problemas mais comuns em organizações que adotam IA: o sistema é implantado para uma finalidade, demonstra valor, e gradualmente passa a ser usado para finalidades diferentes sem que ninguém verifique se continua adequado.

Plano de contingência

Todo sistema falha eventualmente. O que distingue organizações responsáveis das irresponsáveis não é a ausência de falhas – é a existência de plano claro para quando elas ocorrem.

Um plano de contingência para IA deve responder a perguntas concretas: quem é notificado quando o sistema apresenta anomalia? Em que prazo? Com que informações? Existe procedimento para suspender o sistema temporariamente enquanto a anomalia é investigada? Como são revisadas as decisões que o sistema tomou durante o período de falha? Como os afetados são comunicados?

Sem essas respostas definidas previamente, a resposta a um incidente será improvisada – o que é, em geral, mais lenta, mais dispendiosa e mais danosa do que a resposta planejada. A existência de plano de contingência também é, do ponto de vista jurídico, evidência de que a organização antecipou a possibilidade de falha e se preparou para ela – o que é parte do dever de cuidado que o uso de IA em contextos de alto impacto impõe.

O que exigir de um fornecedor de IA

Muitas organizações – incluindo a maior parte das delegacias, tribunais e órgãos de segurança pública que usam IA – não desenvolvem seus sistemas. Compram de fornecedores. Esse modelo não transfere a responsabilidade; distribui-a. E para que a distribuição seja justa e a proteção da organização seja real, o que o fornecedor oferece precisa ser contratualmente especificado.

O contrato com fornecedor de IA deve especificar, no mínimo: documentação completa do sistema, incluindo cartão do modelo com limitações conhecidas e métricas de equidade por grupo; direito de auditoria, incluindo acesso a dados de treino e arquitetura do modelo na medida necessária para verificação de conformidade; obrigações de notificação em caso de incidente de segurança ou degradação de desempenho; responsabilidades claras em caso de dano causado por falha ou viés do sistema; e prazo e condições de suporte e atualização.

Contratos que não especificam essas condições deixam a organização contratante – e os afetados pelas decisões do sistema – sem proteção real. O fornecedor disse que funciona não é defesa jurídica suficiente quando o sistema discrimina, erra ou viola direitos.

CAPÍTULO 8

GOVERNANÇA DE IA: QUEM DECIDE, QUEM CONTROLA, QUEM RESPONDE

Toda organização que usa IA tem, implícita ou explicitamente, uma estrutura de governança. A questão não é se ela existe – é se foi projetada com intenção, ou se simplesmente emergiu de hábitos, pressões e decisões não coordenadas que ninguém realmente assumiu como suas.

Governança implícita tende a ser governança deficiente: ninguém sabe exatamente quem aprovou o uso de determinado sistema, ninguém documenta as razões pelas quais certas salvaguardas foram implementadas e outras não, ninguém é responsabilizado quando o sistema falha porque ninguém assumiu formalmente a responsabilidade de supervisioná-lo. É o ambiente perfeito para que erros se acumulem sem correção e para que, quando o incidente acontece, a responsabilidade se dissolva na complexidade institucional.

Governança explícita – projetada, documentada e com papéis claros – é a alternativa. Este capítulo descreve seus componentes essenciais, com atenção especial ao contexto de segurança pública e investigação criminal.

O que governança de IA precisa resolver

Antes de definir estruturas, é útil ser preciso sobre os problemas que a governança de IA precisa resolver. São, essencialmente, três.

O primeiro é o problema da autorização: quem tem poder para decidir que um sistema de IA será implantado em determinado contexto, com determinado escopo, com determinadas salvaguardas? Essa decisão não pode ser deixada ao arbítrio individual de quem tem acesso técnico ao sistema. Ela precisa de processo deliberado, com análise de riscos, avaliação de conformidade e aprovação de instância competente.

O segundo é o problema da supervisão: quem monitora se o sistema está operando dentro dos parâmetros aprovados, se está produzindo os resultados esperados, se emergiu algum problema que não havia sido antecipado? Essa função não pode ser acumulada com a função de operação – quem usa o sistema cotidianamente não tem a distância necessária para supervisioná-lo com objetividade.

O terceiro é o problema da responsabilização: quando algo dá errado – um erro, uma discriminação, um dano – quem responde? A estrutura de governança precisa criar prestação de contas clara: não para punir arbitrariamente, mas para garantir que erros têm donos, que donos têm incentivos para prevenir erros, e que a responsabilidade é proporcional ao papel de cada ator na cadeia de decisão.

A estrutura de aprovação

Nem todo sistema de IA exige o mesmo nível de escrutínio antes da implantação. Um sistema que sugere respostas automáticas para e-mails internos tem potencial de dano muito menor do que um que recomenda abordagens policiais ou que define

pontuações de risco criminal. A estrutura de aprovação deve ser proporcional ao impacto potencial do sistema sobre direitos fundamentais.

Uma categorização simples e funcional divide os sistemas em três níveis. Sistemas de baixo impacto – que assistem em tarefas internas sem afetar direitos de terceiros – podem ser aprovados por gestores da área com verificação técnica básica. Sistemas de impacto moderado – que afetam serviços prestados ao público, mas de forma reversível e com baixo risco de dano grave – exigem aprovação de nível gerencial com avaliação formal de riscos e auditoria antes da implantação. Sistemas de alto impacto – que embasam decisões sobre liberdade, saúde, crédito, emprego ou outras consequências graves para indivíduos – exigem aprovação de instância superior, avaliação de impacto completa, auditoria externa antes da implantação e revisão periódica obrigatória.

Em organizações de segurança pública, sistemas de reconhecimento facial, perfilamento preditivo e análise de risco de reincidência se enquadram inequivocamente na categoria de alto impacto. A aprovação de qualquer um desses sistemas por decisão unilateral de um gestor de tecnologia, sem processo formal que inclua avaliação jurídica e análise de impacto sobre direitos, é falha de governança – com potenciais consequências jurídicas para a organização e para os indivíduos responsáveis pela decisão.

O comitê de ética de IA

Para sistemas de alto impacto, a análise não deve ser responsabilidade exclusiva de uma única pessoa ou área. A complexidade dos riscos envolvidos – técnicos, jurídicos, éticos, operacionais – exige perspectivas múltiplas e independentes. O instrumento adequado é um comitê de ética ou de governança de IA.

A composição ideal desse comitê combina representantes com competências diferentes e com grau suficiente de independência

entre si. Um representante técnico que conheça o sistema em profundidade suficiente para avaliar suas limitações reais. Um representante jurídico que entenda os princípios constitucionais e regulatórios aplicáveis. Um representante operacional que conheça o contexto real de uso e as pressões práticas que os operadores enfrentarão. E, idealmente, um representante externo ou um ouvidor institucional que traga perspectiva independente, sem o viés de quem tem interesse direto na aprovação do sistema.

O comitê não precisa ser grande, nem precisa se reunir com frequência excessiva. Precisa ter processo claro: como uma análise é solicitada, quais documentos são necessários, quais critérios orientam a decisão, como o resultado é documentado e como é comunicado. E precisa ter autoridade real: suas recomendações de não aprovação ou de implantação condicionada precisam ser respeitadas, sem que a pressão operacional possa simplesmente substituir a análise cuidadosa.

Para órgãos de segurança pública, a existência de comitê desse tipo – ou de estrutura equivalente – é também instrumento de proteção institucional. Quando uma decisão de implantação foi submetida a análise formal, com múltiplas perspectivas documentadas, e ainda assim resultou em problema, a organização pode demonstrar que agiu com diligência. Quando a decisão foi unilateral e apressada, essa demonstração é impossível.

O cartão do modelo como instrumento de governança

A documentação técnica do sistema – o cartão do modelo mencionado no capítulo anterior – é também instrumento central de governança. Ele responde, de forma acessível e padronizada, às perguntas que diferentes atores precisam para exercer suas funções.

O gestor que precisa decidir se aprova o sistema encontra no cartão do modelo as limitações conhecidas, as métricas de desem-