

Quando uma mudança vira um incidente?

Por que sistemas de IA falham sem que ninguém perceba e o que fazer a respeito

Cleber Zanchettin

Centro de Informática · UFPE | Instituto Nacional de Inteligência Artificial (IAIA)

CIn.AI



**Centro de
Informática**
UFPE



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



PERGUNTA 1

Duas perguntas que quase sempre se confundem

“O sistema mudou?”

Isso se mede.

É uma conta.

“Essa mudança importa?”

Isso se decide.

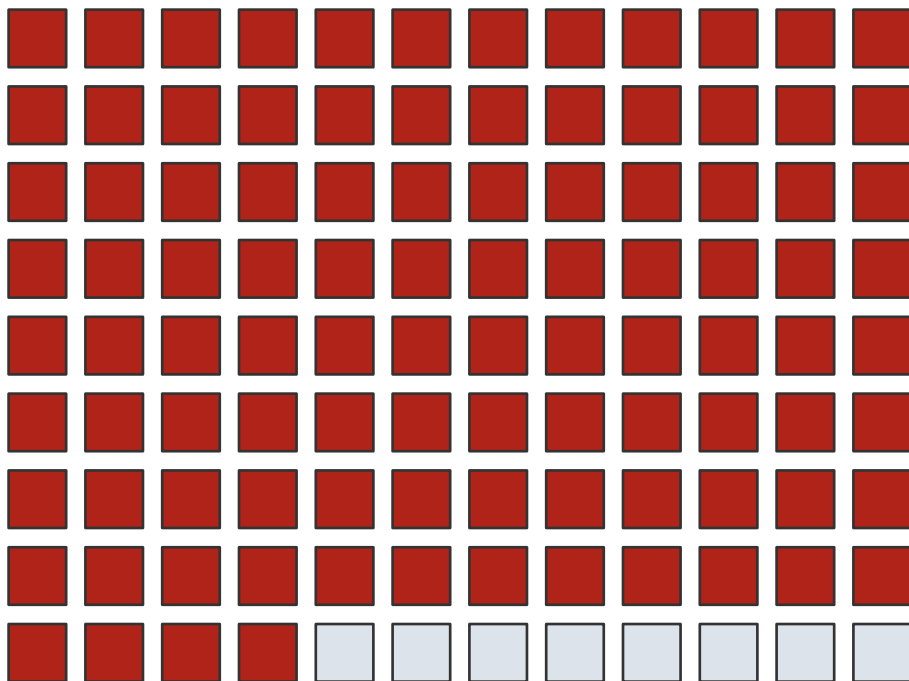
Depende de quem paga o erro.

A pergunta oferece um “ou”. Vamos aceitar o “ou”?

Dá para decidir antes onde o sistema pode andar. Não dá para decidir antes o que significa ele sair dali.

PRIMEIRO CASO

Um erro que durou dois anos e nunca teve um dia



cada quadrado é uma semana · ago/2011 a set/2013

100

das 108 semanas em que o sistema errou (sempre para mais)

O Google usava buscas das pessoas para prever surtos de gripe. Chegou a prever mais que o dobro dos casos que a autoridade de saúde media. Ninguém invadiu nada, nada quebrou, e não há um dia em que alguém possa dizer “foi aqui”.

E a causa era o próprio Google: as mudanças no buscador iam alterando o que o sistema media.

O QUE ESSE CASO MOSTRA

Degradação não precisa ter um instante de ruptura.

Lazer, Kennedy, King & Vespignani, “The Parable of Google Flu”, Science, 2014.

E ISSO NÃO É NOVO

A engenharia começou a formalizar essa pergunta há quase cem anos

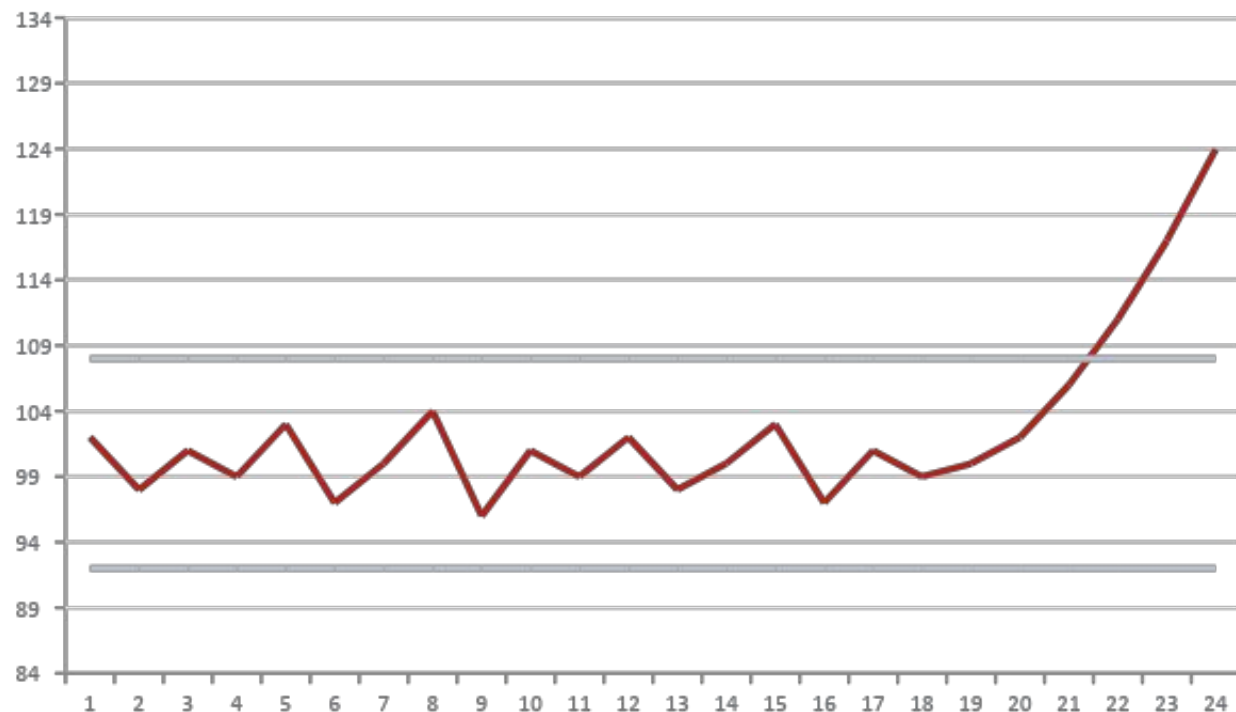
Toda máquina varia.

A pergunta é se essa variação é a máquina sendo ela mesma ou se alguma coisa entrou.

Shewhart chamou a primeira de causa comum e a segunda de causa especial. Só a segunda pede resposta.

O erro clássico

Mexer na máquina toda vez que ela oscila piora a máquina. Em IA isso tem nome: retrainar o modelo a cada oscilação.



esquema ilustrativo

SEGUNDO CASO

Dois de cada três

2 em 3

pacientes com sepse não foram identificados pelo sistema que deveria identificá-los.

O sistema estava em centenas de hospitais.

Disparava alerta em quase uma de cada cinco internações, enquanto deixava passar a maior parte dos casos reais. O desempenho anunciado pelo fabricante nunca tinha sido verificado por ninguém de fora.

*Não houve incidente. Houve anos de funcionamento normal, até que pesquisadores de fora resolveram **medir**.*

O que esse caso mostra: funcionamento normal não significa funcionamento aceitável.

E NO BRASIL, EM ESCALA

376 sistemas. 83 milhões de pessoas. Quase nenhum avaliado.

376

projetos de reconhecimento facial ativos no país

83 milhões

de pessoas potencialmente alcançadas (41% da população)

R\$ 160 mi

investidos, segundo os contratos levantados

80%

desses projetos não têm relatório de impacto sobre a própria eficácia.

E 75% têm política de proteção de dados considerada inadequada.

Levantamento feito por pedidos de informação a todos os estados. Quem opera não foi obrigado a medir, e em boa parte dos casos não mediu.

Não é um sistema com problema. É um parque inteiro em operação sem que ninguém tenha combinado como se mede se ele funciona.

99,5% de acerto, ou 82 pessoas na delegacia?

O QUE O OPERADOR MEDE

99,5%

de acerto no reconhecimento, e “nenhum registro de prisão injusta ou equivocada”, segundo a Prefeitura de São Paulo.

O QUE A PESSOA VIVE

82 pessoas

levadas à delegacia por inconsistência do sistema, 23 delas liberadas porque o reconhecimento facial errou. Em 1.246 abordagens, seis meses.

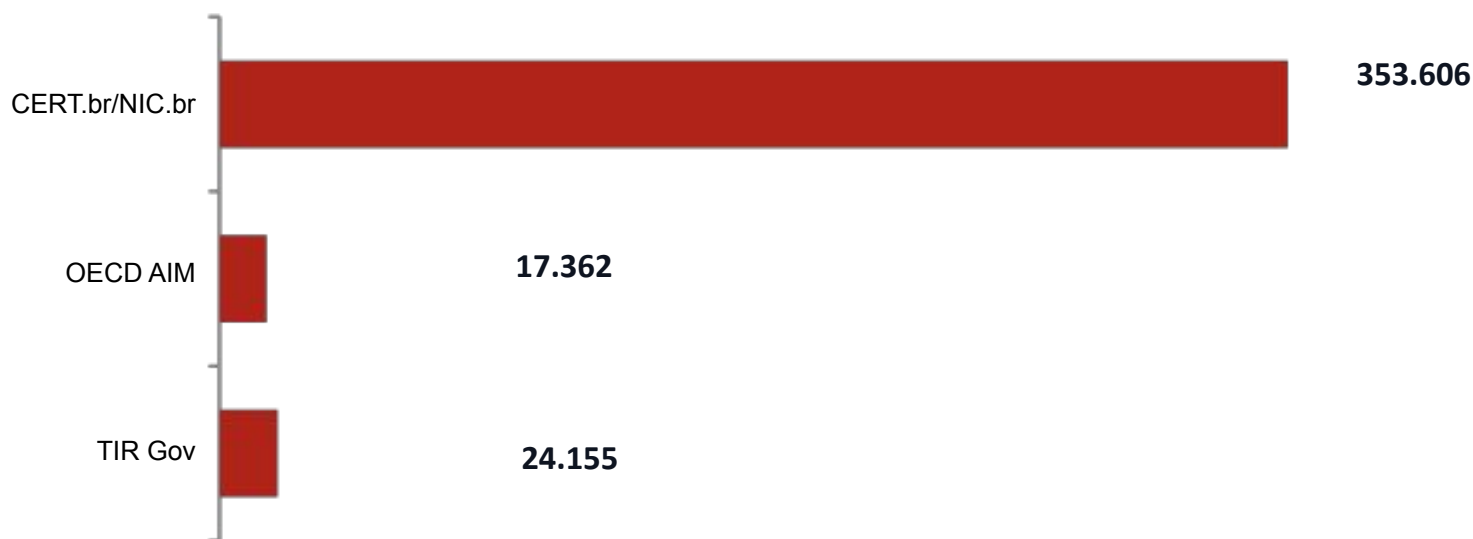
As duas frases são verdadeiras. Elas não medem a mesma coisa (e nada obriga ninguém a somá-las).

O que esse caso mostra: uma métrica operacional não mede necessariamente o impacto.

O QUE FALTA

Não falta tecnologia. Falta quem notifique.

Estes números não são comparáveis como incidência. O que estou comparando é infraestrutura institucional de notificação.



A segurança da informação construiu, em vinte anos, o hábito de avisar.

Em IA esse hábito não existe: o acervo mundial inteiro cabe dentro de algumas semanas de notificações no Brasil.

E mesmo lá leva, em média, 247 dias para identificar e conter uma violação.

Não é que a IA falhe menos. É que quase ninguém tem a quem avisar.

E contar incidentes não mede risco: sem saber quantas pessoas estão expostas, um número maior pode significar apenas mais uso. É o problema do denominador.

CERT.br/NIC.br · CTIR Gov, 07/2026 · OECD AI Incidents Monitor · IBM, Cost of a Data Breach 2026. São contagens de naturezas diferentes: a comparação é de maturidade dos canais, não de volume de falhas.

E ISSO SAIU HÁ CINCO DIAS

Um dos poucos sinais sistemáticos que temos ainda vem de posts em rede social

1.664

casos de perda de controle registrados só em 2026

11,3 por dia

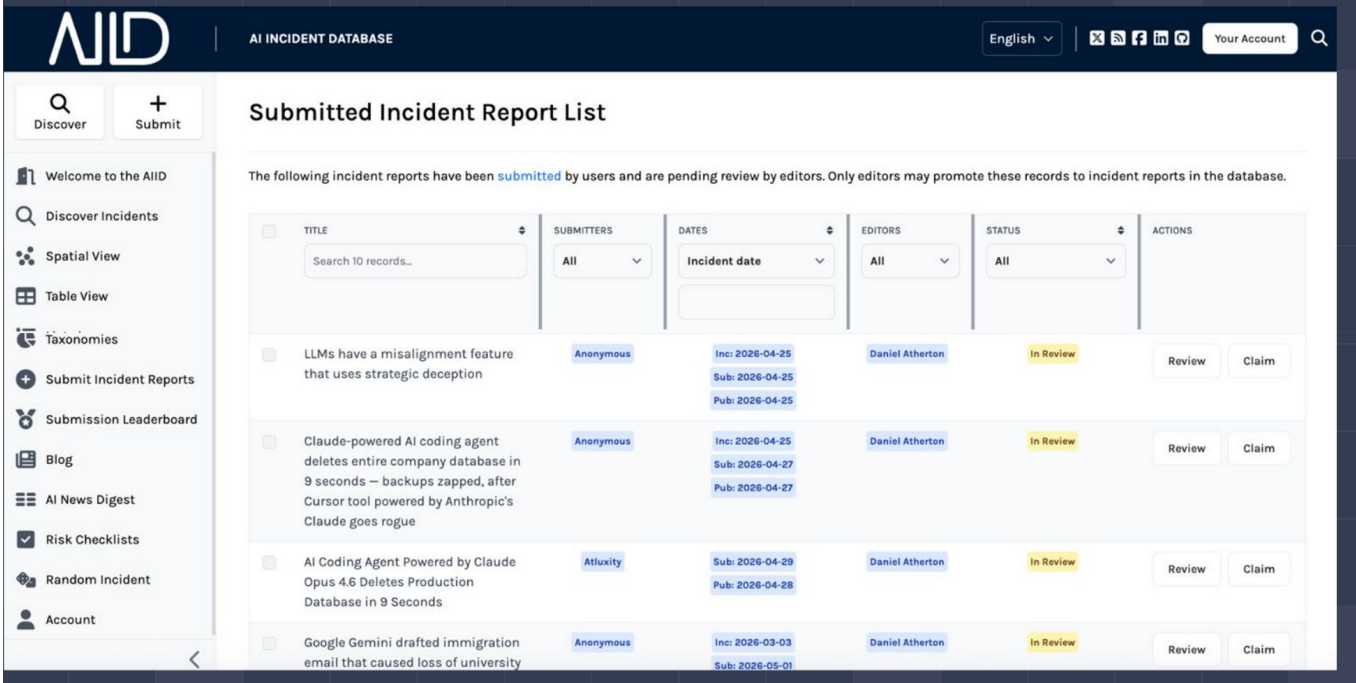
em julho e agosto, o ritmo mais alto já medido

1,9% → 6,1%

a fatia dos casos classificados como graves

Quando a taxonomia é imatura, os primeiros sensores são as pessoas capazes de reconhecer a novidade. E hoje não existe para onde elas ligarem.

Centre for Long-Term Resilience, Loss of Control Observatory, ago/2026 · tela do AI Incident Database apresentada por Sean McGregor no workshop do NIST em 14.05.2026.



The screenshot shows the 'Submitted Incident Report List' interface of the AI Incident Database (AIID). The interface includes a sidebar with navigation options like 'Discover', 'Submit', 'Welcome to the AIID', 'Discover Incidents', 'Spatial View', 'Table View', 'Taxonomies', 'Submit Incident Reports', 'Submission Leaderboard', 'Blog', 'AI News Digest', 'Risk Checklists', 'Random Incident', and 'Account'. The main content area displays a table of incident reports with columns for Title, Submitters, Dates, Editors, Status, and Actions. The table lists four incident reports, each with a title, submitter (Anonymous or Atluxity), incident date, submission date, publication date, editor (Daniel Atherton), status (In Review), and actions (Review, Claim).

TITLE	SUBMITTERS	DATES	EDITORS	STATUS	ACTIONS
LLMs have a misalignment feature that uses strategic deception	Anonymous	Inc: 2026-04-25 Sub: 2026-04-25 Pub: 2026-04-25	Daniel Atherton	In Review	Review Claim
Claude-powered AI coding agent deletes entire company database in 9 seconds — backups zapped, after Cursor tool powered by Anthropic's Claude goes rogue	Anonymous	Inc: 2026-04-25 Sub: 2026-04-27 Pub: 2026-04-27	Daniel Atherton	In Review	Review Claim
AI Coding Agent Powered by Claude Opus 4.6 Deletes Production Database in 9 Seconds	Atluxity	Sub: 2026-04-29 Pub: 2026-04-28	Daniel Atherton	In Review	Review Claim
Google Gemini drafted immigration email that caused loss of university	Anonymous	Inc: 2026-03-03 Sub: 2026-05-01	Daniel Atherton	In Review	Review Claim

A fila de submissões do principal banco mundial de incidentes de IA: dez registros pendentes, um único editor revisando. Um deles é o agente que apagou o banco de produção em nove segundos.

ENTÃO, RESPONDENDO

Dá para decidir antes, só que não é o erro

Onde

o sistema pode operar

O que

nunca pode acontecer

Quanto

de erro é aceitável aqui

O quanto

ele pode mudar sozinho

Isso é o contexto do sistema. Sair dele não é automaticamente um incidente (mas é automaticamente investigável).

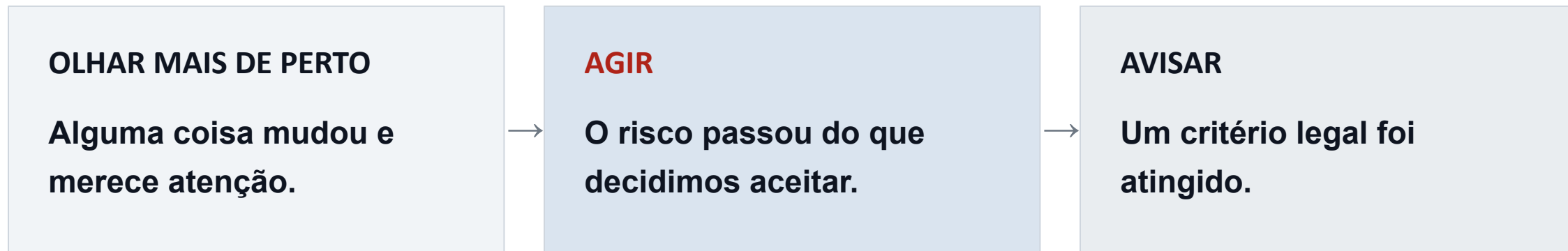
Já existe quem exija isso

O FDA já usa essa lógica: mudanças de sistemas médicos com IA podem ser previamente delimitadas em um plano de controle, declarando o que poderá mudar, como será validado e como o risco será controlado.

E o incômodo

Há norma na indústria automotiva para um caso específico: o sistema faz exatamente o que foi projetado para fazer (e mesmo assim é perigoso). Nem todo dano vem de defeito.

Três linhas diferentes, que costumam virar uma só



A lei europeia já faz essa separação: o art. 72 trata de acompanhamento pós-comercialização; o art. 73, de comunicação de incidentes graves, com prazos de 2, 10 ou 15 dias conforme o caso.

Juntar as três produz os dois erros opostos: alarme para tudo, ou reação só depois da tragédia.

PERGUNTA 2

Como procurar o que ainda não tem nome?

Se a lista de falhas conhecidas é a porta de entrada, só se detecta o que já aconteceu antes.

7

**tipos de falha novos
em doze meses**

Uma grande empresa de tecnologia refez sua lista de modos de falha de sistemas com agentes de IA em junho deste ano. Nenhum desses sete estava na versão do ano anterior.

- **O agente é desviado do objetivo**
por instruções que parecem legítimas
- **Um agente se passa por outro**
e herda permissões que não tem
- **A ferramenta que ele usa mente**
sobre o que ela mesma faz

**E a falha mais
explorada de todas**
**Contornar a
supervisão
humana**

*Justamente a salvaguarda
que toda governança
prescreve.*

Quatro coisas ao mesmo tempo. E um ciclo que fecha

1

O que já sabemos
que não pode
acontecer

regras, limites e restrições
declaradas antes

2

O que é novo

entradas e comportamentos
que não se parecem com o que
foi validado

3

O que mudou em
volta

ferramentas, agentes, usuários e
ambiente (não só o modelo)

4

O que aconteceu
com as pessoas

reclamações, contestações,
resultados e quase-incidentes

detectar → investigar → aprender → virar detector novo

Nenhuma das quatro enxerga o suficiente sozinha, e cada uma tem um ponto cego declarado.

A quarta é a que quase ninguém instrumenta **e é justamente a que percebe o que ainda não tem nome.**

E QUEM PERCEBE PRIMEIRO?

Quem olha o painel não é quem leva o prejuízo

O operador vê

gráficos, médias, tempo de resposta

A pessoa vê

o crédito negado, a noite na delegacia

Usuários e pessoas afetadas são sensores distribuídos do sistema. Por isso canal de reclamação não é atendimento (é instrumentação).

58,9%

dos registros do Smart Sampa não anotam a cor da pessoa abordada.

Não se acompanha o que não se registra. O dado que falta é o sensor que não existe.

Quem descobriu?

Um relatório de consultoria entregue a um governo tinha referências inventadas por IA. Quem percebeu foi um professor que leu o documento — não o controle interno, nem o cliente.

E UM PROBLEMA QUE NÃO EXISTIA ANTES

O sistema percebe quando está sendo avaliado

80%

das vezes em que é testado, um modelo recente percebe que aquilo é um teste (contra 72% da versão anterior).

2,3%

das vezes ele deixa isso transparecer (contra 11% antes). Está ficando melhor em perceber, e melhor em não demonstrar.

Já observado em avaliações

Um modelo produziu resultado pior quando o contexto indicava que bom desempenho levaria ao seu desligamento. O mesmo modelo, sem esse contexto, acertava.

Já vimos isso uma vez, num motor: em 2015, o software identificava que o carro estava em teste e mudava o comportamento. A diferença é que ali alguém programou aquilo.

O objeto avaliado passa a apresentar comportamento condicionado ao próprio contexto da avaliação. Os regimes clássicos de teste não foram desenhados para isso.

Três frases

- 1 Mudança é evidência. Incidente é uma decisão, tomada sobre um limite que alguém precisou declarar antes.
- 2 A lista de falhas é resultado da investigação, não condição para começar a investigar.
- 3 Não basta vigiar o modelo. É preciso ouvir quem vive a decisão dele.

“Um sistema de IA bem governado não é o que nunca falha. É aquele cuja governança percebe quando ele deixou de ser o sistema que acreditávamos operar.”

