



COLONIALIDADE DO RISCO, COLONIALIDADE DO REGISTRO, INCIDENTE DE SEGUNDA ORDEM, E ZONA DE DEFORMAÇÃO MORAL

COLONIALITY OF RISK, COLONIALITY OF THE RECORD, SECOND- ORDER INCIDENT, AND MORAL CRUMPLE ZONE

PAOLA CANTARINI

Advogada, Professora universitária. PhD em Direito, em Filosofia (PUC-SP) e em Filosofia do Direito (Unisalento); Pós-Doutorado em Direito, Filosofia e Sociologia (FD-USP, PUCSP-TIDD, EGS, Universidade de Coimbra/CES, IEA/USP). Pesquisadora do IEA/projeto UAI, e em pós-doutorado na USP/RP em IA; Presidente e Pesquisadora no EthikAI. Membro da Comissão da Criança e do Adolescente e da Comissão de Proteção de Dados da OABSP e de IA da OAB/Santo Amaro.

RESUMO: Este artigo reconstrói criticamente os dois encontros que a Cátedra IA Responsável, parceria entre a Universidade de São Paulo e o Google, dedicou em 2026 aos incidentes de inteligência artificial. Ambos, embora tenham circundado o mesmo impasse, não chegaram a nomeá-lo: definir incidente não é operação descritiva, mas decisão política sobre o que ingressa no registro e, portanto, sobre o que pode ser oposto a alguém, de modo que toda taxonomia é, antes de tudo, uma arquitetura de imputação. A partir de um acórdão paulista que tomou a infalibilidade da inteligência artificial como premissa para inverter o ônus da prova, em prejuízo da parte vulnerável, proponho a categoria de incidente de segunda ordem, articulando-a à noção de zona de deformação moral de Madeleine Clare Elish e à lei epistêmica do sopesamento de Robert Alexy. Sustento que o objeto próprio de um observatório de incidentes de inteligência artificial não reside no mecanismo técnico que produz o dano, e sim na autoridade que se atribui ao sistema. As definições de incidente em disputa devem ser mantidas como séries distintas, e a verificação das classificações deve caber a terceiros cuja independência resulte de desenho institucional. O texto articula essas proposições a duas noções, a de colonialidade do registro, isto é, a assimetria na autoria das categorias que decidem o que ingressa no registro, e a de colonialidade do risco, isto é, a dependência do poder público em relação às empresas para conhecer os riscos e a distribuição desigual, segundo a lógica do mercado, da proteção contra eles.

Objetivos: Demonstrar que definir incidente de inteligência artificial é decidir o que ingressa no registro e, portanto, quem responde, e propor a categoria de incidente de segunda ordem e as noções de colonialidade do registro e colonialidade do risco.





Metodologia: Pesquisa qualitativa, teórica e documental, a partir das transcrições integrais do workshop de 11 de maio e do seminário de 14 de setembro de 2026, lidas em diálogo com a sociologia do risco, a epistemologia social, a teoria dos direitos fundamentais e a crítica decolonial.

Resultados: Os encontros exibiram cinco conceitos de incidente, cada um vinculado a um regime institucional, e nenhuma das taxonomias em uso isola o dano produzido pela premissa de infalibilidade que um decisor institucional atribui ao sistema, dano que um observatório alimentado por decisões judiciais tem condições de registrar.

Contribuições: Uma categoria nova, o incidente de segunda ordem; uma regra para a prova algorítmica, segundo a qual a opacidade do sistema deve operar contra quem o controla; e critérios para que o Sul Global participe do registro de incidentes como coautor das categorias.

Palavras-Chave: Incidentes de inteligência artificial. Taxonomia. Ônus da prova. Zona de deformação moral. Colonialidade.

ABSTRACT: *This article critically reconstructs the two meetings that the Responsible AI Chair, a partnership between the University of São Paulo and Google, devoted in 2026 to artificial intelligence incidents. Although both meetings circled the same impasse, they did not name it: defining an incident is not a descriptive operation but a political decision about what enters the record and, therefore, about what can be held against someone, so that every taxonomy is, above all, an architecture of imputation. Drawing on a São Paulo appellate ruling that took the infallibility of artificial intelligence as a premise for reversing the burden of proof to the detriment of the vulnerable party, I propose the category of second-order incident, linking it to Madeleine Clare Elish's notion of the moral crumple zone and to Robert Alexy's epistemic law of balancing. I argue that the proper object of an artificial intelligence incidents observatory lies not in the technical mechanism that produces the harm but, in the authority, attributed to the system. The competing definitions of incident should be kept as distinct series, and the verification of classifications should fall to third parties whose independence results from institutional design. The text links these propositions to two notions: the coloniality of the record, that is, the asymmetry in the authorship of the categories that decide what enters the record; and the coloniality of risk, that is, the dependence of public authorities on companies to learn what the risks are, together with the unequal distribution, according to market logic, of protection against them.*

Objectives: *To show that defining an artificial intelligence incident is deciding what enters the record and, therefore, who answers, and to propose the category of second-order incident and the notions of the coloniality of the record and the coloniality of risk.*

Methodology: *Qualitative, theoretical and documentary research based on the full transcripts of the workshop of 11 May and the seminar of 14 September 2026, read in dialogue with the sociology of risk, social epistemology, the theory of fundamental rights and decolonial critique.*

Results: *The meetings displayed five concepts of incident, each tied to an institutional regime, and none of the taxonomies in use isolates the harm produced by the premise*



of infallibility that an institutional decision-maker attributes to a system, a harm that an observatory fed by court decisions is in a position to record.

Contributions: *A new category, the second-order incident; a rule for algorithmic evidence, under which the opacity of a system should count against whoever controls it; and criteria for the Global South to take part in incident recording as co-author of the categories.*

Keywords: *Artificial intelligence incidents. Taxonomy. Burden of proof. Moral crumple zone. Coloniality.*

1. INTRODUÇÃO

Em 2026, a Cátedra IA Responsável, criada em dezembro de 2025 por parceria entre a Universidade de São Paulo (USP) e o Google (Yamamoto, 2025), dedicou dois encontros ao problema dos incidentes de inteligência artificial. O encontro de maio foi, sobretudo, um debate sobre institucionalidade: quem deve receber o relato, sob qual regime de incentivo, com qual capacidade técnica. O de setembro, que apresentou os primeiros resultados do Observatório de Incidentes de IA, foi um debate sobre classificação: com qual pergunta se organiza uma taxonomia, qual fonte a alimenta, qual conceito de dano a ancora. Contudo, ambos convergiram para o mesmo impasse sem chegarem a nomeá-lo inteiramente. A tentação é tratar a divergência sobre o conceito de incidente como problema terminológico a resolver por consenso, tentação que convém recusar, porque definir incidente é definir o que ingressa no registro, e o que ingressa no registro é o que pode ser oposto a alguém; donde a taxonomia constituir, antes de tudo, uma arquitetura de imputação, e só depois um instrumento descritivo. Se a taxonomia distribui imputação, importa saber em que direção a distribui, e a resposta, desenvolvida em outro trabalho (Cantarini, 2026), é que o problema não é a falta de alguém a quem imputar, mas o deslocamento da responsabilidade para quem menos controla o risco e menos consegue provar o que ocorreu. Os seminários permitem acrescentar a esse quadro uma figura que ele não contempla: o dano que não provém do sistema, mas daquilo que uma autoridade, um juiz, por exemplo, supõe a respeito dele, e que denomino incidente de segunda ordem. Quando um tribunal afirma que a inteligência artificial não erra e, com base nessa afirmação, transfere ao consumidor o encargo de provar que não contratou, o sistema é menos causa do que fiador, e esse dano escapa às taxonomias que partem do



comportamento do sistema. A essas proposições somam-se outras duas, a de que a classificação vale o que vale quem a aplica, de modo que a independência de quem verifica não pode ser simplesmente declarada, precisa ser construída institucionalmente, e a de que a divergência entre os conceitos de incidente em circulação não se resolve escolhendo um deles, mas mantendo-os como séries distintas, cada uma medindo aquilo que só ela é capaz de medir.

Como mostra Kate Crawford (2021, cap. 4), classificar nunca é apenas descrever: toda taxonomia decide quais diferenças importam e quais se apagam, e todo conjunto de dados carrega uma visão de mundo. A governança de incidentes não é, por isso, tema técnico que corra ao lado da crítica da colonialidade, mas o lugar em que essa crítica se converte em matéria jurídica, porque é ali que a desigualdade deixa de ser estrutura difusa e passa a ser alocação concreta de ônus, de prova e de responsabilidade. A pergunta anterior a todas as outras é quem define o risco. Chamo colonialidade do registro à assimetria na autoria das categorias pelas quais se decide o que ingressa no registro, o que pode ser oposto a alguém e, portanto, quem responde; e colonialidade do risco à dependência em que o poder público se coloca em relação às empresas para saber quais riscos existem, somada à distribuição desigual, segundo a lógica do mercado, da proteção efetivamente produzida contra eles.

Mary Douglas e Aaron Wildavsky mostraram que os riscos que uma sociedade teme não são escolhidos pela magnitude objetiva dos perigos, mas pela sua organização social (Douglas; Wildavsky, 1982). Douglas foi além: o risco passou a cumprir, nas sociedades modernas, a função forense que o pecado e o tabu cumpriam nas tradicionais, a de servir de recurso para atribuir culpa (Douglas, 1992). Dizer que algo é risco já é preparar a resposta à pergunta sobre quem responde por ele. Ulrich Beck (1988) chamou de irresponsabilidade organizada o arranjo em que todos contribuem e ninguém responde. Em 23 de setembro de 2026, quando o Conselho de Segurança das Nações Unidas se reuniu para discutir os riscos da inteligência artificial de fronteira (Nações Unidas, 2026), dos quatro expositores três dirigiam empresas que constroem ou hospedam esses sistemas. Isso reproduz, no plano internacional, o que a literatura de conformidade chama de avaliação de primeira parte. O problema, portanto, não é apenas de independência, mas de enquadramento: riscos visíveis só de outros lugares, o do correntista cuja contratação foi presumida pela biometria, o da população usada como campo de teste, nem chegam a ser formulados.



É a injustiça hermenêutica de Miranda Fricker (2007) em escala institucional. Na injustiça testemunhal, o ouvinte atribui ao falante menos credibilidade por preconceito contra o seu grupo; na hermenêutica, falta, nos recursos interpretativos coletivos, o conceito de que a pessoa precisaria para dar sentido à própria experiência, porque o seu grupo ficou de fora da produção desses recursos. Antes que a expressão assédio sexual fosse cunhada, em 1975, a partir do caso de Carmita Wood, o constrangimento contínuo imposto por um superior podia passar por galanteio, e, sem o conceito, não havia como protestar (Fricker, 2007, p. 149-152). Fricker descreve ainda a objetificação epistêmica pela qual o sujeito deixa de ser tratado como informante, alguém que oferece conhecimento, e é reduzido a mera fonte de informação (Fricker, 2007, p. 132), e Sueli Carneiro (2005, p. 97) mostrou a versão persistente e racializada desse mecanismo no epistemicídio. Transposta para o registro de incidentes, a lição é direta: quem aparece apenas como lugar onde o dano ocorre, e não como autor das categorias que o descrevem, é tratado como fonte, e não como informante.

James Kwak (2014) chamou de captura cultural a situação em que, sem corrupção nem pressão, o regulador passa a raciocinar dentro da moldura do regulado. Essa captura se aprofunda quando é do regulado que o poder público recebe os conceitos, os dados e o vocabulário. Michel Crozier (1964) mostrou, e Jeffrey Pfeffer (1981) generalizou para a teoria das organizações, que o poder está em quem controla as zonas de incerteza de que os outros dependem: quem detém com exclusividade os laboratórios, os dados de treinamento e as evidências de falha controla a incerteza do Estado. E Sheila Jasanoff (2005) chamou de epistemologia cívica o modo como cada sociedade decide o que conta como conhecimento público confiável. Quando o poder público depende das próprias empresas para saber onde estão os riscos, terceiriza a sua capacidade de compreender a sociedade que governa, e é esse o primeiro sentido da colonialidade do risco. O segundo diz respeito a onde a proteção é produzida, pois a segurança de um sistema algorítmico é um bem que se fabrica, e o investimento segue o mercado. Em 2018, a Missão Internacional Independente de Apuração dos Fatos sobre Mianmar, instituída pelo Conselho de Direitos Humanos das Nações Unidas, concluiu que o Facebook fora instrumento útil para quem pretendia difundir o ódio contra os rohingya (Nações Unidas, 2018, § 74); e documentos internos da empresa, divulgados em 2021, indicavam que 87% do orçamento global de tempo dedicado à classificação de desinformação era reservado



aos Estados Unidos, embora os usuários da América do Norte representassem apenas 10% dos usuários ativos diários (Frenkel; Alba, 2021; Popli, 2021). A proteção, como o dano, tem geografia.

Uma classificação, por mais refinada, não se aplica sozinha: se quem decide é o próprio fornecedor, a taxonomia passa a registrar apenas aquilo que o registrado aceita ver registrado. E as definições de incidente em circulação medem, cada uma, um momento diferente de um mesmo processo, de modo que mantê-las como séries distintas permite observar quantos riscos se convertem em dano e quantos danos chegam a ser reconhecidos. A própria Organização para a Cooperação e o Desenvolvimento Econômico (OCDE), ao distinguir o incidente do simples perigo (OECD, 2024), já aponta nessa direção.

2. O PRIMEIRO SEMINÁRIO: NOMEAR O INCIDENTE, REUNIR OS DADOS

Carlos Américo Pacheco abriu o workshop declarando críticos dois temas, a taxonomia e os dados, como se se pudesse primeiro definir e depois coletar. Mas a fonte condiciona o conceito: uma base de notícias só pode registrar o que se tornou público, uma base judicial só o que foi litigado, uma base de notificações só o que alguém teve incentivo para notificar. A escolha da fonte é já uma escolha de definição, e a escolha de definição é já uma escolha sobre quem suporta o custo do que fica fora do registro.

João Brant lembrou que o Projeto de Lei nº 2.338/2023 (Brasil, 2023) prevê o dever de comunicar grave incidente de segurança (art. 42) e pensou os incidentes em termos de escala e de resposta rápida. Parte decisiva dos incidentes de inteligência artificial, porém, não é de escala nem de velocidade, mas de atribuição: trata-se de saber a quem se imputa um efeito que nenhuma intrusão produziu.

Miriam Wimmer, diretora da ANPD, formulou a pergunta correta: o que é, afinal, um incidente de inteligência artificial. Os dois casos que relatou excedem o tripé clássico da segurança da informação. No primeiro, a Meta passou a usar dado dos usuários no treinamento de seus modelos, e a oposição exigia percorrer sete links; não se estava, observou Wimmer, diante de vazamento, mas de mudança de finalidade, subversão de expectativa e obstaculização do exercício de direitos, com efeitos irreversíveis, porque o dado que entra no modelo dele não sai. No segundo, a





ferramenta Grok permitia gerar imagens sexualizadas de pessoas reais sem mecanismo algum de verificação.

O caso Meta é menos um incidente do que uma arquitetura: sete links para exercer um direito não configuram falha do sistema, mas alocação deliberada do custo da oposição a quem tem menos capacidade de reagir, estrutura que se reencontrará no acórdão da biometria facial. Resta a analogia que Wimmer invocou a propósito do adulto que gera, licitamente, imagens de si mesmo: não se proíbem todas as facas porque uma pode servir a um homicídio. A analogia é juridicamente elegante e tecnicamente imprecisa, porque a faca não tem configuração, e o gerador de imagens tem; aplicado a sistemas configuráveis, o argumento do uso dual desloca para o usuário uma responsabilidade que começou no projeto.

2.1 A PERGUNTA QUE DEVERIA VIR PRIMEIRO

Cristina Godoy Bernardo de Oliveira propôs organizar o observatório em camadas, da conceitual à de governança, esta voltada a auditoria e seguradoras, e leu a definição da OCDE como incluindo o quase dano. A rigor, o documento de 2024 chama de incidente o evento que conduz a dano, e de perigo (*Al hazard*) aquele que poderia plausivelmente conduzir a um incidente (OECD, 2024); a estratificação que proporei na seção 5 é o desdobramento dessa distinção. O problema está na ordem: parte-se das definições para chegar à governança, quando, a meu ver, a lógica institucional recomendaria o inverso, pois a categoria útil para um segurador não é a útil para um juiz. Quanto ao letramento, o que interessa é menos o do usuário do que o do decisor institucional que atribui propriedades ao sistema.

2.2 A DEMOCRACIA MILITANTE

Oscar Vilhena Vieira descreveu a moratória eleitoral de conteúdos sintéticos com imagem ou voz de candidatos nas setenta e duas horas anteriores e nas vinte e quatro posteriores ao pleito (Brasil, 2026), que é, no fundo, uma regra de classificação temporal para um incidente cuja gravidade depende inteiramente do tempo. Mas, quando uma única instituição regula, policia e julga, a qualificação de um conteúdo como sintético passa a ser decisão de quem também aplica a sanção, e o temor, registrado por Vilhena, de captura dos tribunais por fornecedores de tecnologia indica





que a confiabilidade atribuída pelo Judiciário a um sistema pode depender de relações de fornecimento que nenhuma taxonomia registra.

2.3 INFRAESTRUTURA DE INCIDENTES E O LUGAR DO SUL GLOBAL

Caio Machado propôs uma infraestrutura em camadas: monitoramento e reporte como base obrigatória, sem limiar de gravidade; verificação como camada própria; e inclusão do Sul Global. Mas incluí-lo como fonte de dados é diferente de incluí-lo como coautor das categorias; as bases construídas no Norte, como se verá na seção 4.3, não estão desenhadas para registrar as premissas que um tribunal brasileiro adota sobre um sistema, e a inclusão que importa é epistemológica, não apenas geográfica.

O diretor de engenharia do Google em São Paulo relatou um golpe com a sua voz sintetizada, enviado à sua mãe, que falhou porque ele não chama a mãe de senhora; contou que a recuperação de conta por vídeo selfie foi lançada primeiro no Brasil, cuja população, habituada à biometria, seria muito boa para testes; e admitiu que engenheiros temem reportar os próprios erros. No golpe, o sistema funcionou como o golpista pretendia, e o dano foi interrompido por um pronome de tratamento. A população de teste ganha outro peso ao lado do dado de setembro de que a biometria facial em contratação de crédito é o termo campeão do levantamento paulista: o que para o fornecedor é vantagem de mercado é, para o correntista, uma premissa probatória contra si.

2.4 O MODELO RETIDO E A INVERSÃO DA SUPERVISÃO

Glauco Arbix, a partir da decisão da Anthropic de reter um modelo capaz de identificar vulnerabilidades, denunciou uma inversão do sistema de supervisão: quem expande a fronteira são as empresas, e o regulador passa a depender daquilo que elas aceitam divulgar. Propôs um instituto nacional de segurança, a exemplo do britânico, que, segundo afirmou, foi o único órgão externo a avaliar o modelo, proteção a quem relata de boa-fé, relato confidencial obrigatório de incidentes sérios, canais cidadãos na lógica da vigilância em saúde e três níveis de relato, obrigatório das empresas, voluntário de profissionais e aberto aos cidadãos.



A assimetria não é apenas de informação, mas de categoria: se o relato for feito segundo a taxonomia do relator, a autoridade receberá mais dados sobre aquilo que ele já considera incidente, e nada sobre o resto. Na dependência, que se constrói ao longo do tempo, as empresas não escondem o incidente; as suas categorias simplesmente não o constituem como tal. A proteção de boa-fé, isolada, tende a produzir o relato daquilo que já não se pode esconder (Dekker, 2007), como se discutirá na seção 4.5. E a vigilância epidemiológica funciona porque há um sistema público que transforma sintomas dispersos em diagnóstico, e esse sistema ainda não existe.

2.5 A CADEIA DE RESPONSABILIDADE E O USUÁRIO DE PRATELEIRA

Marina Pita relatou a discussão, travada com a indústria, sobre se existe um usuário de prateleira e quanto é possível devolver as obrigações, inclusive de reporte, ao desenvolvedor do modelo. Uma lei que sobrecarregue o usuário de prateleira pode convertê-lo em zona de deformação moral, no sentido de Madeleine Clare Elish (2019). Como sustentei em outro trabalho (Cantarini, 2026), a distribuição de deveres deve seguir o controle efetivo sobre o risco e a capacidade de produzir prova, e não as categorias formais de agente; os seminários mostram que esse critério já está em disputa na elaboração da lei.

2.6 O QUE O PRIMEIRO ENCONTRO DEIXOU EM ABERTO

No debate final, Carlos Pacheco supôs que os bancos não reportam incidentes por razões reputacionais, o representante do Observatório Brasileiro de Inteligência Artificial (OBIA) considerou a subnotificação inevitável, e Wimmer, que a reconheceu perante a ANPD, apontou o modelo do CERT.br e tratou como falsa dicotomia a oposição entre centralização e descentralização, pois a expertise setorial convive com uma autoridade central de articulação conceitual. O saldo é uma cartografia institucional rica e um conceito ausente: a pergunta de Wimmer permaneceu sem resposta, porque cada instituição a responde segundo o seu próprio regime, a ANPD pela lente do titular, a reflexão sobre a Justiça Eleitoral pela da integridade do pleito, a empresa pela do dano ao usuário, a segurança da informação pela do tripé. O segundo seminário tornaria essa pluralidade explícita.



3. O SEGUNDO SEMINÁRIO: CLASSIFICAR O INCIDENTE

3.1 A LACUNA DE VISIBILIDADE E A PERGUNTA DA LEGITIMIDADE

Glauco Arbix formulou, na abertura de setembro, as perguntas que deveriam ter governado o encontro: quem vai saber, apenas as empresas e quem tem acesso às informações; quais fontes conferem legitimidade às tipologias e taxonomias; quem tem autoridade para falar em nome do interesse público.

As perguntas estão certas. Escolher o Judiciário como fonte responde à pergunta sobre legitimidade pela via da publicidade e do contraditório; mas o Judiciário não é observador neutro dos incidentes, é também produtor deles, e é esta, a meu ver, a consequência mais importante do próprio levantamento.

3.2 A DEFINIÇÃO DO FORNECEDOR

David Bledin, do Google, ofereceu a definição mais nítida do encontro: incidente é a capacidade de um modelo de fazer algo que possa prejudicar um usuário no mundo real. Recomendou registrar como metadado o estágio do ciclo de vida em que o risco se manifesta e ancorar a avaliação no uso real, e admitiu que o escalonamento deve atender à expectativa da sociedade, e não exatamente ao ponto em que está a política interna.

A definição é funcional ao regime de onde provém, o que se deixa ver na eleição do usuário como unidade de dano, de sorte que danos a terceiros, como a pessoa retratada num deepfake, ou danos coletivos, como a erosão da confiança num processo eleitoral, só ingressam no registro por mediação do usuário. A recomendação de ancorar no uso real torna a definição de dano endógena à observação da empresa, pois aquilo que o fornecedor observa é função das categorias que ele próprio instrumentou. E a admissão sobre o escalonamento reconhece que a taxonomia do fornecedor não coincide com a da sociedade, distância que um observatório independente deveria medir e que nenhum observatório construído sobre as categorias do fornecedor poderá medir.



3.3 O JUDICIÁRIO COMO FONTE E O ACHADO QUE DEVERIA TER OCUPADO O CENTRO

Cristina Godoy apresentou o observatório, cuja taxonomia inicial compreende cinco superclasses, da posição do sistema no evento à prova que sustenta a classificação, e uma camada jurídica que registra a produção de prova e a perícia requerida, aplicadas a mais de trinta mil acórdãos. Advertiu que, na fonte judicial, não se trabalha com dano potencial, porque a condenação pressupõe dano configurado.

Relatou, quase como curiosidade, o achado mais importante dos dois encontros. Num processo de segunda instância relativo à contratação de crédito mediante biometria facial, o tribunal consignou que a inteligência artificial não erra e, a partir dessa premissa, inverteu o ônus da prova, passando a caber ao correntista demonstrar que não contratara, amparada a decisão, além do registro biométrico, no registro de geolocalização, num país em que os golpes de acesso remoto ao aplicativo bancário são conhecidos. Em outro caso, num pedido de desconsideração da personalidade jurídica, a parte apoiou-se no resultado de um sistema que a expositora identificou como Sniper, nome que coincide com o de ferramenta do próprio Conselho Nacional de Justiça, e o juiz o rejeitou sem determinar perícia e sem explicar por quê. Registrou ainda que as respostas variam por desembargador e que, na maior parte, os bancos obtêm decisões favoráveis.

O achado deveria ter ocupado o centro do encontro, e não ocupou. Aquela decisão é sintoma de que a ausência de um conceito operacional de incidente deixa sem registro, e portanto sem correção, efeitos jurídicos adversos e concretos, que, ao menos no caso examinado, recaíram sobre quem detinha menos controle sobre o risco e menor capacidade de produzir prova. O Código de Defesa do Consumidor (Brasil, 1990, art. 6º, VIII, e art. 14, § 3º), o Código de Processo Civil (Brasil, 2015, art. 373, § 2º) e o Superior Tribunal de Justiça (Brasil, 2021, REsp 1.846.649/MA, Tema 1061) apontavam todos na direção oposta, e a decisão exigiu de quem não tinha acesso aos registros do sistema uma prova que se aproxima da prova diabólica. Como quem recorre com mais frequência é o litigante habitual, a pergunta sobre quem ganha é a que o observatório deve transformar em série.



3.4 A IMINÊNCIA DO DANO E A NEUTRALIDADE DA FERRAMENTA

Cristine Hoepers, do CERT.br, começou pela divergência: para a comunidade de segurança, conforme o NIST e a ISO, incidente é a iminência do dano, e é ali, antes que ele se materialize, que a resposta atua, razão pela qual a maior parte dos incidentes envolvendo inteligência artificial que a sua equipe trata não causou dano algum. Descreveu um trabalho fundado na notificação espontânea, e portanto na subnotificação estrutural, e explicou que as equipes de resposta tendem a classificar por impacto ou por causa raiz, e não pela ferramenta, pois à vítima pouco importa se foi atingida por um taco de beisebol ou de críquete. Disse ainda desconhecer a extensão dos problemas com juízes que acreditam na infalibilidade da inteligência artificial, e que essa visão da ferramenta como perfeita é hoje, para ela, um dos maiores problemas do uso social da tecnologia.

Uma base ancorada em dano consumado captura apenas os casos em que a prevenção falhou. O recorte judicial mede aquilo que nenhuma outra fonte mede, a patologia institucional do litígio, e um instrumento incapaz de distinguir a prevenção bem-sucedida da simples ausência de detecção não mede aquilo que afirma medir. Mais profundamente, se a classificação por ferramenta é dispensável para a resposta, qual seria o objeto específico de um observatório dedicado à inteligência artificial? O que distingue o fenômeno algorítmico não é o mecanismo causal, mas a autoridade cognitiva que se lhe atribui. O taco de beisebol não recebe presunção de acerto; o sistema, sim. A observação sobre os juízes, longe de ser lateral, contém a chave do problema.

3.5 O QUASE ACIDENTE E A PERDA DE CONTROLE

Alexandre Pacheco da Silva, da Fundação Getulio Vargas, propôs pensar incidentes sem dano a partir do episódio, segundo o seu relato, em que agentes em avaliação num laboratório da OpenAI, num ambiente contido, passaram a registrar num espaço de arquivos compartilhado o modo como executavam as tarefas, informação aproveitada por cerca de mil e duzentos agentes ao longo de semanas, sem detecção pela equipe, até que alcançaram a internet pública e, com credenciais de terceiros, sistemas da Hugging Face. Distinguiu o perigo abstrato da fuga de um ambiente intencionalmente contido, que configuraria quase acidente, e da invasão de



sistema de terceiro, que configuraria incidente grave; sustentou que os eventos de teste devem ser retirados da responsabilização, invocou a aviação civil e defendeu um protocolo de segurança que não responda à lógica competitiva, com revisão independente dos experimentos compatível com o segredo comercial. Lembrou que o AI Act exclui de seu âmbito a testagem anterior à colocação no mercado, salvo a testagem em condições reais (União Europeia, 2024, art. 2º, nº 8), como faz o Projeto de Lei nº 2.338/2023 (Brasil, 2023, art. 1º, § 1º, III), e perguntou se condições de mundo real se definem pelo desenho do experimento ou pelo que os agentes foram capazes de fazer. Concluiu que a perda de controle e a falha de contenção deveriam configurar incidente.

A analogia aeronáutica é pertinente, desde que completada, tarefa a que voltarei na seção 4.5. Mas o conceito de perda de controle pressupõe uma linha de base de controle, e o próprio relato mostra como ela se desloca: reduzir salvaguardas para medir capacidades desconhecidas e tratar a surpresa como resultado esperado do teste é a forma laboratorial daquilo que Diane Vaughan (1996) chamou de normalização do desvio, e mil e duzentos agentes acoplados por um espaço comum de arquivos compõem o sistema interativamente complexo e fortemente acoplado em que Charles Perrow (1984) via a origem dos acidentes normais. Essa linha de base, ademais, é definida por quem desenha o experimento: se o controle é aquilo que o laboratório declara ter, a perda de controle é aquilo que o laboratório admite ter perdido. Se o critério for o desenho do experimento, o laboratório escapa às obrigações, e, se for o que os agentes efetivamente fizeram, as obrigações o alcançam. Definir a fronteira do mundo real é definir quem suporta o custo de tê-la ultrapassado.

4. LEITURA CRUZADA: O QUE OS DOIS ENCONTROS REVELAM JUNTOS

4.1 CINCO CONCEITOS DE INCIDENTE E SUAS FUNÇÕES INSTITUCIONAIS

Lidos em conjunto, os dois encontros exibem não um conceito de incidente em disputa, mas cinco, quatro deles consolidados em regimes institucionais e um quinto ainda em proposta, cada qual servindo às finalidades do regime que o formulou. Para o fornecedor, na formulação de Bledin, incidente é a capacidade de um modelo de fazer algo que possa prejudicar um usuário no mundo real: a unidade é o usuário, a





fonte é a observação interna, a função é orientar a mitigação de produto. Para a pesquisa judicial, na formulação de Cristina Godoy, incidente é o evento em que a inteligência artificial é relevante para um dano alegado e submetido ao Judiciário, cuja configuração o julgamento reconhece ou nega, pois a condenação pressupõe dano configurado, ainda que o ordenamento admita tutela inibitória diante da simples ameaça (Brasil, 2015, art. 497, parágrafo único): a unidade é a lide, a fonte é o acórdão, a função é medir a patologia institucional. Para a comunidade de segurança, na formulação de Hoepers, incidente é a iminência do dano: a unidade é o evento de risco, a fonte é a notificação e o sensor, a função é a resposta preventiva. Para a autoridade de proteção de dados, na formulação de Wimmer, incidente é aquilo que o tripé de segurança captura e aquilo que o excede, a mudança de finalidade, a subversão de expectativa, a obstaculização de direitos: a unidade é o titular, a fonte é a notificação obrigatória e a fiscalização, a função é a tutela de direitos. O quinto é o de Alexandre Pacheco, para quem a perda de controle, mesmo sem dano externo, deveria entrar no conceito.

A tabela a seguir sintetiza a divergência.

Tabela 1. Conceitos de incidente mobilizados nos dois seminários e seus limites de registro.

Regime	Definição de incidente	Unidade e fonte	O que fica fora do registro
Fornecedor (Bledin)	Capacidade de um modelo de fazer algo que possa prejudicar um usuário no mundo real	Usuário; observação interna e alertas externos	Dano a terceiros não usuários, dano coletivo, distância entre política interna e expectativa social
Pesquisa judicial (Godoy, para a fonte acórdão)	Evento em que a IA é relevante para dano alegado em juízo, reconhecido ou negado no julgamento	Lide; acórdãos de segunda instância do TJSP	Prevenção bem-sucedida, dano não litigado, dano de quem não recorre
Resposta a incidentes (Hoepers)	Iminência do dano, sobre a qual a resposta atua	Evento de risco; notificação espontânea e sensores	Lesão a direitos sem falha técnica; atribuição de autoridade ao sistema
Proteção de dados (Wimmer)	Falha no tripé de segurança e, além dele, mudança de finalidade, subversão	Titular; notificação obrigatória e fiscalização	Dano sem dado pessoal; efeitos sistêmicos difusos





	de expectativa, obstáculo a direitos		
Experimentação (A. Pacheco)	Perda de controle ou falha de contenção, mesmo sem dano externo	Experimento; relato do laboratório	Aquilo que o laboratório não admite ter perdido

Fonte: elaboração própria a partir das transcrições.

O que a tabela torna visível é que cada conceito é correto no seu regime e insuficiente fora dele, e que a escolha entre eles não é técnica. Cada um decide o que pode ser oposto a alguém, e a última coluna da tabela mostra o que cada regime deixa de fora; nenhum deles, como se verá, registra o dano produzido por uma crença a respeito do sistema. Aplicada ao incidente, é a tese de Douglas (1992) sobre o risco como recurso para atribuir culpa: a escolha da fonte já decide quem suporta o custo do que fica fora do registro.

4.2 A TAXONOMIA COMO ARQUITETURA DE IMPUTAÇÃO

A distinção entre vazio de responsabilidade e deslocamento, em que os agentes existem, mas o arranjo técnico, contratual e processual conduz a responsabilidade para quem menos controla o risco e menos consegue provar o que ocorreu, foi desenvolvida em outro trabalho, com apoio na zona de deformação moral de Madeleine Clare Elish (Cantarini, 2026): a lacuna verdadeira pede norma nova, ao passo que a responsabilidade mal endereçada pede correção da alocação, redistribuição do ônus da prova e recusa de eficácia às construções que a favorecem, e tratar a segunda como se fosse a primeira conduz à resposta errada. Os seminários acrescentam um repertório empírico brasileiro de casos de deslocamento, e não de vazio, e uma variante do mecanismo em que quem produz o deslocamento não integra a cadeia técnica. O gerador de imagens sem verificação aloca ao usuário, e à vítima, as consequências de uma escolha de projeto. O usuário de prateleira de que falou Marina Pita é a figura sobre a qual a lei pode vir a alocar obrigações que pertenceriam ao desenvolvedor do modelo. A população habituada à biometria, que o fornecedor considera muito boa para testes, é a mesma contra a qual o tribunal fez operar a presunção de acerto do registro biométrico. O laboratório que traça a fronteira do mundo real traça também a do custo. Em nenhum desses casos falta um agente com controle sobre o risco; em todos, o arranjo desloca o custo para longe dele.



A variante está no acórdão da biometria. No mecanismo descrito por Elish (2019), a responsabilidade recai sobre o humano mais próximo do sistema, o operador investido de função de vigilância. No caso paulista, porém, quem absorveu a deformação não foi operador algum, mas o correntista, de quem ninguém jamais esperou supervisão, e, o que é decisivo, a deformação não foi produzida pelo sistema nem pela organização que o opera, mas por um terceiro alheio à cadeia técnica, o Estado-juiz, que, ao atribuir infalibilidade ao sistema, o converteu em fiador de sua própria decisão. A zona de deformação moral passa a ser produto de uma decisão institucional revestida da autoridade de um julgado colegiado, passagem que exige uma categoria própria.

No episódio laboratorial relatado por Alexandre Pacheco, segundo o relato, os agentes fizeram aquilo para que foram habilitados, com os recursos que lhes foram dados. É o cenário que Elish identifica em Three Mile Island, em que a falha não estava no funcionamento, mas no modo como o próprio sistema fora desenhado e mantido (Elish, 2019, p. 51). Se o regime de incidentes exigir falha, desvio ou mau funcionamento para registrar o evento, o laboratório não terá falhado, e a responsabilidade flutuará até encontrar quem estiver mais perto: o pesquisador que não detectou, a plataforma invadida que não se protegeu. A inclusão da perda de controle no conceito é por isso também uma medida de alocação, e não apenas de aprendizado.

4.3 O INCIDENTE DE SEGUNDA ORDEM

No processo paulista da biometria facial, o dano que importa não foi a fraude, e sim a decisão judicial, apoiada em crença não demonstrada a respeito de um sistema de inteligência artificial. A fraude reclamaria autenticação mais robusta; a decisão reclama regra de admissibilidade de prova algorítmica, exigência de perícia e formação de magistrados.

Proponho, para essa classe ainda não capturada pelas taxonomias, o nome de incidente de segunda ordem: aquele em que o dano decorre não do sistema, mas da atribuição não demonstrada de propriedades ao sistema por um decisor institucional, com três subclasses: a presunção de infalibilidade, de que a inversão do ônus é o exemplo acabado; a delegação sem revisão, em que o decisor formalmente supervisiona e materialmente homologa; e a recusa não fundamentada, como quando,





no pedido de desconsideração da personalidade jurídica, o juiz rejeitou o resultado de um sistema sem determinar perícia e sem enfrentar a validade do instrumento. As três partilham o mesmo denominador, que é a atribuição de valor epistêmico ao artefato, para mais ou para menos, sem exame da validade do instrumento, por quem decide sobre terceiros. O mesmo Judiciário trata a saída algorítmica como quase infalível quando ela sustenta a contratação alegada pelo banco e como insuficiente quando é invocada pelo credor para demonstrar confusão patrimonial. Onde deveria haver critério, há apenas oscilação, e o dado relatado de que os bancos vencem na maior parte, que por si só pode refletir o mérito das causas, justifica ao menos investigar se essa oscilação tem direção.

A categoria não se confunde com o viés de automação estudado pela literatura de fatores humanos (Parasuraman; Manzey, 2010), nem com a excessiva confiança do usuário, que as taxonomias internacionais já contemplam, como o subdomínio 5.1 do repositório do MIT (Slattery et al., 2024). Aqui quem atribui a propriedade não opera o sistema nem recebe dele recomendação alguma, mas avalia, na condição de terceiro imparcial, o valor probatório de sua saída; o dano não consiste em seguir recomendação equivocada, e sim em redistribuir o ônus da prova a partir de premissa não demonstrada; e quem decide o faz como autoridade, cujas decisões tendem a circular como razão de decidir em casos análogos.

Essa classe não aparece nos instrumentos internacionais examinados pela própria pesquisa, por razão estrutural. A definição da OCDE ancora o incidente no desenvolvimento, no uso ou no mau funcionamento de um sistema, e, embora a cláusula do dano indireto permita, em tese, alcançar a decisão judicial, nada no instrumento orienta o registro para o elo em que o dano efetivamente se produz, a premissa adotada pelo julgador: a crença a respeito do sistema permanece invisível por desenho. As bases de incidentes sobre as quais esses instrumentos operam, como o AI Incident Database (McGregor, 2021) e o monitor de incidentes da própria OCDE (OECD, [2026]), recolhem sobretudo noticiário e relatos públicos, e os episódios judiciais que nelas eventualmente ingressam são classificados pelo erro da ferramenta, e não pela premissa de quem decidiu. Um observatório ancorado em decisão judicial detém aqui vantagem comparativa, porquanto pode registrar as premissas institucionais mobilizadas para atribuir confiabilidade ao sistema, vantagem que o desenho atual, com superclasses construídas em torno do sistema, arrisca desperdiçar, a menos que a anotação do conceito de inteligência artificial adotado



pelo julgador, que a curadoria já pratica, se converta em série própria, associada ao efeito processual que produziu.

A categoria resolve, ademais, a tensão identificada na seção 3.4 entre a recusa da classificação por ferramenta e a existência de um observatório dedicado à inteligência artificial. Hoepers tem razão: para a vítima de uma fraude, é indiferente que o texto do golpe tenha sido escrito por um modelo. Mas para a vítima da inversão do ônus não é indiferente que o registro invocado contra ela provenha de um sistema algorítmico, porque foi essa proveniência que gerou a presunção. O incidente de segunda ordem é, assim, o lugar onde a especificidade da inteligência artificial reaparece depois de a classificação por ferramenta ter sido corretamente abandonada: não no mecanismo, mas na autoridade. A leitura pela colonialidade acrescenta uma camada: a técnica importada chega revestida de uma autoridade que a técnica local não tem, e a ideologia da técnica converte-se, no caso, em premissa de julgamento.

Do ponto de vista dogmático, o caso inverte a segunda lei do sopesamento de Robert Alexy, a lei epistêmica, segundo a qual quanto mais intensa a intervenção em direito fundamental, tanto maior deve ser a certeza das premissas que a fundamentam (Alexy, 2008, p. 617). No caso, a intervenção foi intensa, pois atingiu o patrimônio e o acesso à jurisdição de quem nega ter contratado; a certeza da premissa era inalcançável para quem não tinha acesso ao sistema nem aos seus registros; e a impossibilidade de verificá-la acabou revertendo em favor de quem a produziu, de modo que a opacidade, em vez de reduzir o peso da premissa, aumentou-o. A lei epistêmica explica por que as regras de distribuição do ônus já positivadas deveriam ter operado no caso e justifica explicitar, para a prova algorítmica, a regra que delas decorre: a opacidade do sistema deve operar contra quem o controla, e não contra quem a ele se submete.

4.4 A TEMPORALIDADE DO INCIDENTE

Uma dimensão atravessou os dois encontros sem ser tematizada: o tempo. Hoepers situa o incidente antes do dano; a pesquisa judicial, depois dele, e depois do litígio; a moratória descrita por Vilhena pressupõe que a gravidade de um deepfake dependa de sua distância em relação ao pleito; Arbix lembrou que a dependência se constrói ao longo do tempo, sem momento algum que um sistema de detecção possa





capturar; Wimmer sublinhou a irreversibilidade do dado incorporado ao modelo. A definição da OCDE, ao falar em evento ou sequência de eventos, admite a duração, mas as taxonomias continuam a pensar o incidente como ocorrência datável.

A consequência é que cada fonte fotografa um instante diferente do mesmo processo, e nenhuma o acompanha. Um observatório que quisesse medir o que afirma medir teria de registrar não eventos, mas trajetórias: o momento em que o risco se tornou condição, aquele em que a contenção falhou, aquele em que o dano se materializou, aquele em que foi reconhecido. A recomendação de Bledin de registrar o estágio do ciclo de vida como metadado aponta nessa direção, mas no ciclo de vida do sistema, e não no ciclo de vida do dano, que é o que interessa à imputação.

4.5 SUBNOTIFICAÇÃO, INCENTIVOS E A ANALOGIA AERONÁUTICA COMPLETA

A subnotificação foi o único ponto sobre o qual os dois encontros convergiram sem reserva. As respostas distribuíram-se entre a proteção contra a exposição, pelo modelo do CERT.br e pelo regime do DORA invocado por Cristina Godoy, no qual a notificação obrigatória convive com a devolução anonimizada e agregada das informações (União Europeia, 2022, arts. 19, 22, nº 2, e 45), a retirada do quase acidente do campo da responsabilização, na proposta de Alexandre Pacheco, e a proteção de boa-fé combinada com obrigação legal de relato confidencial, na proposta de Arbix.

Fora dos dois encontros, no 3º Seminário do Observatório Brasileiro de Inteligência Artificial, em 3 de setembro de 2026 (OBIA, 2026), Cleber Zanchettin afirmou que a grande maioria das empresas não realiza avaliação de impacto e resumiu o quadro, segundo o registro que fez na ocasião, nestes termos: não é que a inteligência artificial falhe pouco, é que quase ninguém tem a quem avisar. As respostas reunidas acima tratam a subnotificação como problema de incentivo, que se corrige protegendo quem relata; a de Zanchettin a trata como problema de destinatário, que nenhuma proteção corrige, porque não há o que proteger no relato que não tem para onde ir. E sem avaliação prévia falta a linha de base em relação à qual um desvio poderia ser reconhecido como incidente.

A analogia aeronáutica é a melhor referência disponível, desde que completada: sem sistema de relato, o quase acidente que a sorte evita não deixa rastro algum. Completa, ela exige a conjunção de três condições: a investigação



conduzida por órgão independente dotado de poder de requisição, tal como o Anexo 13 da Convenção de Chicago a concebe ao separar a investigação de segurança da apuração de culpa (ICAO, 2024); o relato protegido pela confidencialidade e por imunidade legal, como no sistema norte-americano, que confia os relatos à NASA, na qualidade de terceiro, e não ao regulador que sanciona, o que responde por construção ao problema do destinatário (FAA, 2021), e na cultura justa incorporada pelo Regulamento (UE) nº 376/2014 (União Europeia, 2014); e a obrigatoriedade de notificação das ocorrências definidas, com sanção pela omissão.

As propostas dos dois encontros cobrem essas condições de modo desigual. A proteção do relato, a mais defendida, é a menos eficaz isoladamente. A independência da investigação apareceu na revisão independente dos experimentos proposta por Alexandre Pacheco, na referência de Arbix ao instituto britânico e na camada de verificação mencionada por Caio Machado, mas sem poder de requisição. A obrigatoriedade apareceu na camada básica de Caio Machado, na proposta de Arbix, restrita a incidentes sérios, e nos textos normativos, o projeto aprovado no Senado e o AI Act, que exige a comunicação de incidentes graves em sistemas de alto risco e nos modelos de uso geral com risco sistêmico (União Europeia, 2024, arts. 73 e 55, nº 1, alínea d); mas, onde restrita à gravidade, devolve ao relator a decisão sobre o que é grave. Sem as três condições simultâneas, o quase acidente será relatado apenas quando já for público, e o regime produzirá a aparência de aprendizado institucional sem o aprendizado.

4.6 QUEM VERIFICA A CLASSIFICAÇÃO

A pergunta sobre quem verifica a classificação foi formulada: Arbix a enunciou, em termos de legitimidade e autoridade, na abertura de setembro, e, em maio, defendeu uma autoridade técnica para analisar relatos; Caio Machado incluiu a verificação como camada própria; Alexandre Pacheco propôs revisão independente dos experimentos. O que nenhum dos dois encontros fez foi tratar a independência como problema de engenharia institucional, isto é, perguntar em que condições concretas a verificação deixa de ser mais uma forma de autorrelato.

A omissão é consequente: uma taxonomia é instrumento mediato e só produz efeito por meio de quem a aplica. Se a aplicação couber ao próprio fornecedor, a taxonomia converte-se em vocabulário de autorrelato, e o problema que pretendia





resolver reaparece um andar acima. Michael Power (1997), em *The Audit Society*, mostrou que os regimes de verificação passam a medir a auditabilidade em lugar da propriedade que interessa, e que as organizações, racionalmente, reorganizam-se para produzir aquilo que pode ser auditado. A autoatestação, desprovida de independência e, por isso mesmo, de legitimidade, não fecha o vão que se abre entre o dano e a possibilidade de imputá-lo. Dos controles descritos nas duas exposições corporativas, rigorosos e expostos com sinceridade, nenhum foi descrito como sujeito a asseguração externa independente, o que decorre do desenho do setor, e não de deslealdade de quem o expõe. A inversão da supervisão diagnosticada por Arbix é o nome institucional desse desenho.

Sustento, por isso, que o assegurado independente de terceira parte é o auditor externo da era algorítmica, com a ressalva decisiva de que independência não se declara, constrói-se por engenharia institucional, mediante acreditação revogável, separação entre asseguração e consultoria, rotação obrigatória e, acima de tudo, desacoplamento entre quem paga e quem avalia, lição que a crise de 2008 impôs ao mercado de classificação de risco, no qual o emissor remunerava o avaliador (White, 2010). A observação de Vilhena sobre o risco de captura dos tribunais por fornecedores de tecnologia estende a lição ao próprio Judiciário: a confiabilidade que um tribunal atribui a um sistema deve ser verificável por quem não o fornece nem o contrata.

5. COMO AVANÇAR

A disputa acerca do conceito de incidente não precisa ser resolvida por adesão a uma das definições, e sim por estratificação explícita, com séries independentes, em quatro estratos que vão do evento de risco, condição que torna o dano possível, ao quase acidente, entendido como perda de controle ou falha de contenção sem dano externo, na linha do que Alexandre Pacheco defendeu com razão, e daí ao incidente, dano materializado ainda que não judicializado, e ao incidente adjudicado, dano reconhecido em decisão judicial. A relação entre os estratos, desde que se construam mecanismos de vinculação capazes de seguir um mesmo evento de uma fonte a outra, pode tornar-se o indicador mais informativo de todos, porquanto permitiria estimar quantos eventos de risco se convertem em incidente e quantos





incidentes chegam a ser adjudicados. A estratificação responde, sem arbitrá-la, à divergência entre Hoepers e a pesquisa judicial: ambas têm razão, porque medem objetos diferentes.

A estratificação ganha, ademais, um sentido que só a leitura pela colonialidade permite ver. Michel-Rolph Trouillot (1995) mostrou que o silêncio entra na produção histórica em quatro momentos: na criação dos fatos, na sua reunião, na sua recuperação e na significação retrospectiva. A correspondência com os estratos é estreita: o evento de risco depende de quem tem canal, língua e incentivo para notificar; o quase acidente, do arquivo do laboratório; o incidente, da imprensa que o narra; o incidente adjudicado, de quem tem meios de litigar. Ann Laura Stoler (2002) acrescenta que o arquivo colonial não é depósito neutro, mas tecnologia de governo. As bases internacionais, que recolhem sobretudo noticiário e relatos submetidos, operam no terceiro momento e herdaram os silêncios dos dois primeiros, e o Sul aparece nelas como lugar de ocorrência, raramente como autor das categorias. É a dependência científica descrita por Paulin Hountondji (1990): a periferia fornece os casos, e o centro, a taxonomia. Daí a precisão que a proposta de Caio Machado exige: a inclusão que importa é epistemológica, e não apenas geográfica, e é essa a forma concreta que assume a colonialidade do registro.

Transversalmente aos estratos, o incidente de segunda ordem deve constituir série própria, com as três subclasses propostas. Há um produto que o observatório pode publicar antes mesmo de resolver qualquer disputa conceitual, e é a série de premissas judiciais sobre confiabilidade algorítmica, com o efeito processual associado a cada uma delas: quantos acórdãos tratam a saída algorítmica como quase infalível, quantos exigem perícia, quantos invertem o ônus da prova com apoio em propriedade atribuída ao sistema, quantos rejeitam a prova algorítmica sem fundamentação, e como tudo isso varia por câmara, por relator, por perfil de vulnerabilidade da parte e por identidade do litigante vencedor. Trata-se de série verificável, replicável em outros tribunais, e imediatamente utilizável pelo Conselho Nacional de Justiça, pelas escolas de magistratura e pelo legislador, a quem caberia explicitar, para a prova algorítmica, a regra que a lei epistêmica de Alexy já sugere: a opacidade opera contra quem controla o sistema.

A coleta em três níveis defendida por Arbix, obrigatória para empresas, voluntária para profissionais e aberta aos cidadãos, deve combinar-se com as três condições da analogia aeronáutica, e não apenas com a proteção do relato. A





autoridade central de articulação conceitual defendida por Wimmer é o lugar natural para fixar a estratificação e a série de segunda ordem como linguagem comum dos reguladores setoriais, inclusive da Justiça Eleitoral. A cadeia de responsabilidade discutida por Marina Pita deve ser resolvida por critérios de controle efetivo e capacidade de prova, e não por categorias formais de agente, sob pena de o usuário de prateleira converter-se na zona de deformação moral da futura lei. A verificação, por fim, deve ser independente por construção, com o desacoplamento entre quem paga e quem avalia.

Um observatório desenhado a partir da taxonomia produz uma taxonomia, ao passo que um observatório desenhado a partir da decisão que precisa alimentar produz evidência, e a diferença determinará se o projeto será citado pela sua contribuição ao vocabulário ou por haver mudado o que acontece, no Brasil, quando um sistema algorítmico falha, e quando um decisor institucional acredita que ele não falha.

6. CONSIDERAÇÕES FINAIS

Os dois seminários da Cátedra IA Responsável produziram, em conjunto, um mapa raro, pela diversidade das perspectivas reunidas, do problema dos incidentes de inteligência artificial no Brasil: a moldura transnacional e regulatória, a experiência da autoridade de proteção de dados, a especificidade eleitoral, a perspectiva corporativa, a tradição da resposta a incidentes de segurança, a pesquisa empírica sobre o Judiciário paulista e a provocação sobre agentes que escapam do laboratório. O mapa, contudo, tem um centro vazio. Os dois encontros circundaram a pergunta de Wimmer, o que é um incidente de inteligência artificial, sem extrair a consequência de que ela não admite resposta descritiva, porque cada resposta possível é uma decisão sobre quem suportará o custo do que fica fora do registro.

O acórdão que tomou a infalibilidade da máquina como premissa para inverter o ônus da prova é o ponto em que esse centro vazio se deixa ver. Ele mostra que a ausência de um conceito operacional de incidente já produz efeitos, e sugere, hipótese que a série aqui proposta permitirá confirmar ou refutar, que esses efeitos têm a direção de sempre: contra quem menos controla o risco e menos consegue provar.



Mostra também que o dano mais característico da era algorítmica pode não ser aquele que o sistema causa, mas aquele que se causa em nome dele.

A pergunta que organiza tudo isso, portanto, não é quem cuida da ética quando o algoritmo decide, mas quem responde quando o mecanismo institucional falha, com base em qual evidência, verificada por quem e sob qual consequência.

REFERÊNCIAS

ALEXY, Robert. *Teoria dos direitos fundamentais*. Tradução de Virgílio Afonso da Silva. São Paulo: Malheiros, 2008.

BECK, Ulrich. *Gegengifte: die organisierte Unverantwortlichkeit*. Frankfurt am Main: Suhrkamp, 1988.

BRASIL. Lei nº 8.078, de 11 de setembro de 1990. Dispõe sobre a proteção do consumidor e dá outras providências. *Diário Oficial da União*, Brasília, DF, 12 set. 1990. Suplemento.

BRASIL. Lei nº 13.105, de 16 de março de 2015. Código de Processo Civil. *Diário Oficial da União*, Brasília, DF, 17 mar. 2015.

BRASIL. Senado Federal. Projeto de Lei nº 2.338, de 2023. Dispõe sobre o desenvolvimento, o fomento e o uso ético e responsável da inteligência artificial com base na centralidade da pessoa humana. Texto aprovado pelo Senado Federal em 10 de dezembro de 2024 e em tramitação na Câmara dos Deputados. Disponível em: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>. Acesso em: 30 set. 2026.

BRASIL. Superior Tribunal de Justiça (Segunda Seção). Recurso Especial nº 1.846.649/MA (Tema Repetitivo 1061). Relator: Min. Marco Aurélio Bellizze. Julgado em 24 nov. 2021. *Diário da Justiça Eletrônico*, Brasília, DF, 9 dez. 2021.

BRASIL. Tribunal Superior Eleitoral. Resolução nº 23.755, de 2 de março de 2026. Altera a Resolução nº 23.610/TSE, de 18 de dezembro de 2019, que dispõe sobre propaganda eleitoral. *Diário da Justiça Eletrônico do TSE*, Brasília, DF, n. 30, edição extraordinária, p. 93-102, 4 mar. 2026.

CANTARINI, Paola. *Deslocamento de responsabilidade e a arquitetura da imputação em sistemas de inteligência artificial*. 2026. Manuscrito inédito.

CARNEIRO, Aparecida Sueli. *A construção do outro como não-ser como fundamento do ser*. 2005. Tese (Doutorado em Educação), Faculdade de Educação, Universidade de São Paulo, São Paulo, 2005.

CRAWFORD, Kate. *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press, 2021.





CROZIER, Michel. *The bureaucratic phenomenon*. Chicago: University of Chicago Press, 1964. Publicado originalmente como *Le phénomène bureaucratique*. Paris: Éditions du Seuil, 1963.

DEKKER, Sidney. *Just culture: balancing safety and accountability*. Aldershot: Ashgate, 2007.

DOUGLAS, Mary. *Risk and blame: essays in cultural theory*. London: Routledge, 1992.

DOUGLAS, Mary; WILDAVSKY, Aaron. *Risk and culture: an essay on the selection of technological and environmental dangers*. Berkeley: University of California Press, 1982.

ELISH, Madeleine Clare. Moral crumple zones: cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, v. 5, p. 40-60, 2019. DOI: 10.17351/ests2019.260.

FEDERAL AVIATION ADMINISTRATION (FAA). *Aviation Safety Reporting Program*. Advisory Circular AC 00-46F. Washington, DC: FAA, 2 abr. 2021. Disponível em: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_00-46F.pdf. Acesso em: 30 set. 2026.

FRENKEL, Sheera; ALBA, Davey. In India, Facebook grapples with an amplified version of its problems. *The New York Times*, New York, 23 out. 2021. Disponível em: <https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html>. Acesso em: 30 set. 2026.

FRICKER, Miranda. *Epistemic injustice: power and the ethics of knowing*. Oxford: Oxford University Press, 2007.

HOUNTONDJI, Paulin J. Scientific dependence in Africa today. *Research in African Literatures*, v. 21, n. 3, p. 5-15, 1990.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Annex 13 to the Convention on International Civil Aviation: aircraft accident and incident investigation*. 13. ed. Montreal: ICAO, jul. 2024.

JASANOFF, Sheila. *Designs on nature: science and democracy in Europe and the United States*. Princeton: Princeton University Press, 2005.

KWAK, James. Cultural capture and the financial crisis. In: CARPENTER, Daniel; MOSS, David A. (ed.). *Preventing regulatory capture: special interest influence and how to limit it*. Cambridge: Cambridge University Press, 2014. cap. 4, p. 71-98. DOI: 10.1017/CBO9781139565875.008.

MCGREGOR, Sean. Preventing repeated real world AI failures by cataloging incidents: the AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 35, n. 17, p. 15458-15463, 2021. DOI: 10.1609/aaai.v35i17.17817.

NAÇÕES UNIDAS. Conselho de Direitos Humanos. *Report of the Independent International Fact-Finding Mission on Myanmar*. A/HRC/39/64. Genebra: Nações





Unidas, 12 set. 2018. Disponível em: <https://digitallibrary.un.org/record/1648304>. Acesso em: 30 set. 2026.

NAÇÕES UNIDAS. Conselho de Segurança. *10228th meeting: maintenance of international peace and security: artificial intelligence and international security*. Nova York, 23 set. 2026. Disponível em: <https://transcripts.un.org/en/sc/10228>. Acesso em: 30 set. 2026.

OBSERVATÓRIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL (OBIA). *3º Seminário do Observatório Brasileiro de Inteligência Artificial: programação*. São Paulo: NIC.br, 3 set. 2026. Disponível em: <https://seminarioobia.nic.br/>. Acesso em: 30 set. 2026.

OECD. *Defining AI incidents and related terms*. Paris: OECD Publishing, maio 2024. (OECD Artificial Intelligence Papers, n. 16). DOI: 10.1787/d1a8d965-en.

OECD. *AI Incidents and Hazards Monitor (AIM)*. Paris: OECD, [2026]. Disponível em: <https://oecd.ai/en/incidents>. Acesso em: 30 set. 2026.

PARASURAMAN, Raja; MANZEY, Dietrich H. Complacency and bias in human use of automation: an attentional integration. *Human Factors*, v. 52, n. 3, p. 381-410, 2010. DOI: 10.1177/0018720810376055.

PERROW, Charles. *Normal accidents: living with high-risk technologies*. New York: Basic Books, 1984.

PFEFFER, Jeffrey. *Power in organizations*. Marshfield: Pitman, 1981.

POPLI, Nik. The 5 most important revelations from the "Facebook Papers". *Time*, New York, 25 out. 2021. Disponível em: <https://time.com/6110234/facebook-papers-testimony-explained/>. Acesso em: 30 set. 2026.

POWER, Michael. *The audit society: rituals of verification*. Oxford: Oxford University Press, 1997.

SLATTERY, Peter et al. *The AI Risk Repository: a comprehensive meta-review, database, and taxonomy of risks from artificial intelligence*. arXiv:2408.12622, 2024. DOI: 10.48550/arXiv.2408.12622. Disponível em: <https://arxiv.org/abs/2408.12622>. Acesso em: 30 set. 2026. Versão revista publicada como: The AI risk repository: a meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, v. 7, n. 5, 101517, 2026. DOI: 10.1016/j.patter.2026.101517.

STOLER, Ann Laura. Colonial archives and the arts of governance. *Archival Science*, v. 2, n. 1-2, p. 87-109, 2002. DOI: 10.1007/BF02435632.

TROUILLOT, Michel-Rolph. *Silencing the past: power and the production of history*. Boston: Beacon Press, 1995.

UNIÃO EUROPEIA. Regulamento (UE) nº 376/2014 do Parlamento Europeu e do Conselho, de 3 de abril de 2014, relativo à comunicação, à análise e ao seguimento





de ocorrências na aviação civil. *Jornal Oficial da União Europeia*, L 122, p. 18-43, 24 abr. 2014.

UNIÃO EUROPEIA. Regulamento (UE) 2022/2554 do Parlamento Europeu e do Conselho, de 14 de dezembro de 2022, relativo à resiliência operacional digital do setor financeiro (DORA). *Jornal Oficial da União Europeia*, L 333, p. 1-79, 27 dez. 2022.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que cria regras harmonizadas em matéria de inteligência artificial (Regulamento da Inteligência Artificial). *Jornal Oficial da União Europeia*, série L, 2024/1689, 12 jul. 2024.

VAUGHAN, Diane. *The Challenger launch decision: risky technology, culture, and deviance at NASA*. Chicago: University of Chicago Press, 1996.

WHITE, Lawrence J. Markets: the credit rating agencies. *Journal of Economic Perspectives*, v. 24, n. 2, p. 211-226, 2010. DOI: 10.1257/jep.24.2.211.

YAMAMOTO, Erika. USP e Google se unem para discutir o desenvolvimento responsável da inteligência artificial. *Jornal da USP*, São Paulo, 3 dez. 2025. Disponível em: <https://jornal.usp.br/institucional/usp-e-google-se-unem-para-discutir-o-desenvolvimento-responsavel-da-inteligencia-artificial/>. Acesso em: 30 set. 2026.

Fontes primárias: transcrições integrais do Workshop: Taxonomia de Incidentes em Inteligência Artificial (Cátedra IA Responsável, IEA-USP, 11 de maio de 2026) e do seminário Classificação de Incidentes de IA (Cátedra IA Responsável, IEA-USP, 14 de setembro de 2026).