A inteligência artificial (IA) tem avançado a passos largos, transformando setores como saúde, educação e, especialmente, o direito. Modelos de IA generativa, como o ChatGPT, têm sido empregados para auxiliar na elaboração de documentos jurídicos, pesquisa de jurisprudência e até mesmo na interação com clientes. No entanto, um fenômeno preocupante tem emergido: as "alucinações" da IA. Essas alucinações ocorrem quando o sistema gera informações que, embora pareçam plausíveis, são incorretas ou inexistem na realidade. No contexto jurídico, isso pode resultar em interpretações errôneas da lei, citações fictícias ou até mesmo a criação de precedentes inexistentes.

Imagine um advogado utilizando uma ferramenta de IA para pesquisar jurisprudência e, ao receber uma citação de um tribunal que nunca proferiu tal decisão, basear sua argumentação nela. Ou considere um juiz que, ao consultar um assistente virtual para esclarecer um ponto legal, recebe uma explicação imprecisa que influencia sua sentença. Esses cenários ilustram os riscos das alucinações da IA no ambiente jurídico, onde a precisão e a confiabilidade das informações são essenciais.

Além dos riscos jurídicos, as alucinações da IA levantam questões éticas significativas. A confiança excessiva em sistemas automatizados pode levar à disseminação de informações errôneas, comprometendo a integridade do processo judicial e a confiança pública no sistema legal. A falta de regulamentação específica sobre o uso de IA no direito agrava esses desafios, tornando urgente a discussão sobre a responsabilidade civil em casos de falhas da IA.

Este artigo busca explorar as implicações legais das alucinações em IA generativa, apresentando casos reais que evidenciam os riscos e desafios enfrentados pelo setor jurídico. Além disso, serão discutidas possíveis soluções e regulamentações necessárias para mitigar esses problemas, garantindo que a integração da IA no direito seja feita de forma segura e ética.

A compreensão profunda desse fenômeno é crucial para profissionais do direito, desenvolvedores de IA e legisladores, a fim de criar um ambiente jurídico que aproveite os benefícios da tecnologia sem comprometer a justiça e a equidade.

1.O Fenômeno das Alucinações de IA: Causas e Características

O advento de modelos de linguagem natural (MLNs) como Deepseek, Gemini, Chat GPT tem revolucionado diversas áreas, incluindo o setor jurídico. No entanto, a crescente utilização dessas ferramentas traz consigo um desafio significativo: as alucinações. Alucinações em IA referem-se à geração de informações incorretas, inventadas ou enganosas por modelos de IA, apresentadas de forma convincente como se fossem factuais. Este fenômeno, intrínseco ao funcionamento dos MLNs, demanda uma análise aprofundada de suas causas, características e implicações, especialmente em contextos pelas quais a precisão da informação é crucial. Essa necessidade de escrutínio se intensifica ao considerarmos que a IA, em sua essência, opera como um sistema de aproximações e probabilidades, distante da certeza e da veracidade absoluta que se espera em domínios como o direito (Marcus & Davis, 2022).

A raiz das alucinações reside na própria arquitetura e processo de treinamento dos MLNs. Esses modelos são treinados em vastos conjuntos de dados textuais, extraídos da internet, livros, artigos científicos e outras fontes. Durante o treinamento, eles aprendem a identificar padrões e associações estatísticas entre palavras e frases (Bender et al., 2021).

No entanto, essa aprendizagem é puramente estatística e não implica uma compreensão semântica do conteúdo. Os modelos não "entendem" o significado das palavras ou a veracidade das informações; eles apenas preveem a probabilidade de uma sequência de palavras ser "coerente" com base nos dados de treinamento. Nesse sentido, a crítica de Chomsky (1957) à abordagem puramente estatística da linguagem, embora anterior à era da IA moderna, ressoa ao alertar para as limitações de modelos que negligenciam a estrutura sintática profunda e o significado semântico.

Vários fatores contribuem para as alucinações. Um deles é a tendência dos MLNs em extrapolar padrões aprendidos a partir dos dados de treinamento para gerar novas informações. Essa extrapolação, embora fundamental para a capacidade de generalização dos modelos, pode levar à criação de informações que não são precisas ou que não existem na realidade (Shaip, 2022). Este processo de extrapolação é inerente ao design dos modelos, que visa generalizar o conhecimento a partir dos dados observados. No entanto, a generalização excessiva pode resultar em associações espúrias e, consequentemente, em alucinações, como demonstrado por Huang et al. (2021) em seu estudo sobre a geração de notícias falsas por modelos de linguagem. Essa problemática da generalização excessiva é também abordada por Domingos (2015), que argumenta que a busca por "algoritmos mestres" capazes de aprender qualquer coisa a partir de dados pode levar a modelos com baixa capacidade de discriminação e propensos a erros.

A qualidade, diversidade e representatividade dos dados de treinamento também exercem um papel crucial no desempenho dos MLNs. Dados tendenciosos, incompletos ou desatualizados podem levar os modelos a gerar informações falsas ou enganosas. A falta de diversidade nos dados de treinamento pode levar a modelos que refletem apenas uma visão parcial da realidade, exacerbando o risco de alucinações (Gebru et al., 2018). Essa questão da parcialidade dos dados é central no debate sobre a ética da IA, com O'Neil (2016) alertando para os perigos dos "algoritmos de destruição", que perpetuam e amplificam desigualdades sociais.

Outro fator relevante é a falta de consciência contextual dos MLNs. Eles não possuem uma "consciência" do mundo real ou do contexto mais amplo das informações que geram. Eles apenas correlacionam padrões presentes nos dados de treinamento, sem levar em consideração a veracidade ou relevância das informações. Essa falta de consciência contextual pode levar os modelos a "preencher lacunas" com informações que não têm relação com a realidade (Shaip, 2022). Essa limitação dos modelos de linguagem em compreender o contexto é criticada por Searle (1980) em seu famoso argumento do "quarto chinês", que demonstra que a manipulação de símbolos por um sistema, por mais sofisticada que seja, não implica necessariamente uma compreensão genuína do significado.

Modelos de linguagem são frequentemente otimizados para a fluidez e naturalidade da linguagem gerada, o que pode levar a priorizar a coerência linguística em detrimento da precisão factual (Ji et al., 2023). Essa otimização pode resultar na criação de narrativas convincentes, mas factualmente incorretas. Bowman (2015) argumenta que essa ênfase na fluidez pode levar a modelos que reproduzem padrões linguísticos superficiais sem realmente compreender o conteúdo.

Uma característica marcante das alucinações é que, apesar de serem incorretas, as respostas geradas frequentemente parecem plausíveis para os usuários. A fluidez e

naturalidade da linguagem gerada podem levar os usuários a acreditar na veracidade do conteúdo, mesmo quando este é falso ou inventado (Marcus & Davis, 2022). Essa característica torna as alucinações particularmente perigosas em áreas como o direito, onde a precisão das informações é fundamental. Essa plausibilidade enganosa é um dos principais desafios na detecção de alucinações, pois exige uma avaliação crítica e a verificação independente das informações fornecidas pelos modelos.

O setor jurídico, com sua dependência da precisão, precedentes e interpretações complexas, é particularmente vulnerável aos riscos das alucinações em IA. Casos recentes demonstram as consequências potenciais do uso inadequado dessas ferramentas. Advogados nos Estados Unidos e no Brasil foram sancionados por citar casos fictícios gerados por IA em documentos judiciais (Reuters, 2025; Migalhas, 2025). Esses incidentes destacam a necessidade de verificação rigorosa das informações geradas por IA antes de sua utilização em contextos legais. Um advogado em Melbourne foi encaminhado ao órgão de reclamações legais após admitir o uso de software de IA que gerou citações de casos falsos em um tribunal de família (The Guardian, 2024). Esse caso ressalta a importância da supervisão humana e da validação das informações geradas por IA, mesmo quando a fonte parece confiável.

A mitigação das alucinações é um desafio complexo que requer uma abordagem multifacetada. A seleção e curadoria cuidadosa dos dados de treinamento são essenciais para garantir que os modelos de IA tenham uma compreensão equilibrada e abrangente do assunto (Gebru et al., 2018). Isso inclui a remoção de dados tendenciosos, incompletos ou desatualizados, bem como a inclusão de fontes diversas e representativas. No entanto, a simples melhoria dos dados de treinamento pode não ser suficiente para eliminar completamente as alucinações, como argumenta Mitchell (2019), que defende a necessidade de uma abordagem mais fundamental para a construção de modelos de IA que sejam capazes de raciocinar e compreender o mundo de forma mais semelhante aos humanos.

As métricas tradicionais de avaliação de modelos de linguagem, como perplexidade e BLEU, não são adequadas para detectar alucinações. É necessário desenvolver métricas mais robustas que avaliem a precisão factual das informações geradas pelos modelos (Ji et al., 2023). Além disso, é possível incorporar mecanismos de verificação da verdade nos modelos de IA, como a consulta a bases de dados externas ou a utilização de técnicas de raciocínio lógico para validar as informações geradas (Thorne et al., 2018). No entanto, a implementação desses mecanismos de verificação da verdade pode ser complexa e exigir um grande volume de recursos computacionais, como aponta Chollet (2021), que defende a necessidade de uma abordagem mais eficiente e escalável para a avaliação da precisão factual dos modelos de linguagem.

A supervisão humana e a validação rigorosa das informações geradas por IA são essenciais para garantir a precisão e confiabilidade dos resultados, especialmente em áreas críticas como o direito. A busca por modelos de IA mais transparentes e explicáveis pode ajudar a identificar as causas das alucinações e a desenvolver estratégias para mitigá-las (Rudin, 2019). Contudo, a busca por explicabilidade na IA pode ser um desafio complexo, como argumenta Lipton (2018), que aponta para a existência de um "trade-off" entre a precisão e a explicabilidade dos modelos, o que significa que modelos mais precisos tendem a ser menos explicáveis e vice-versa.

Um exemplo notório ocorreu nos Estados Unidos, onde advogados envolvidos em um processo contra o Walmart foram multados por citar casos fictícios gerados por IA em documentos judiciais. O juiz federal Kelly Rankin determinou que os advogados tinham a obrigação ética de verificar a autenticidade dos casos citados, destacando o risco crescente de litígios associados ao uso de IA. Um dos advogados, Rudwin Ayala, reconheceu ter utilizado um programa de IA interno que gerou os casos falsos e foi multado em $3.000. Além disso, Ayala foi removido do caso. Os outros dois advogados, T. Michael Morgan e Taly Goody, receberam multas de $1.000 cada um por não verificar adequadamente a precisão do documento apresentado (Reuters, 2025).

No Brasil, especificamente em Florianópolis, um advogado foi multado por utilizar jurisprudência falsa gerada por IA em um recurso. O profissional admitiu ter utilizado o ChatGPT para elaborar o recurso, justificando que o erro foi um "uso inadvertido" da tecnologia. Apesar da justificativa, a câmara considerou a conduta grave o suficiente para aplicar a multa e encaminhar o caso à Ordem dos Advogados do Brasil de Santa Catarina (OAB/SC) (Migalhas, 2025).

Além disso, um advogado de Melbourne foi encaminhado ao órgão de reclamações legais de Victoria após admitir o uso de software de IA que gerou citações de casos falsos em um tribunal de família, resultando no adiamento de uma audiência. O advogado forneceu ao tribunal uma lista de casos anteriores solicitada pela juíza Amanda Humphreys, mas nem ela nem seus associados puderam identificar esses casos, pois eram inventados pelo software de IA. O advogado admitiu não ter verificado a precisão da informação antes de apresentá-la ao tribunal e ofereceu uma desculpa incondicional, além de pagar os custos da audiência falhada (The Guardian, 2024).

Em um estudo realizado por Marcus e Davis (2022), foi observado que, em modelos de linguagem avançados, a IA não só cria falhas factuais, mas pode gerar informações completamente inventadas, como citações de jurisprudência que não existem ou referências a autores que nunca publicaram sobre o tema em questão. Esse fenômeno pode ter implicações drásticas em contextos jurídicos e acadêmicos, onde a confiabilidade das fontes e a verificação rigorosa são exigidas.

Além disso, outro fator que contribui para as alucinações é a falta de uma "consciência" no modelo de IA sobre o contexto mais amplo das informações que gera. A IA não possui conhecimento do mundo real, mas apenas correlaciona padrões presentes nos dados nos quais foi treinada. Isso significa que, em cenários complexos, como o direito digital ou a interpretação de normas jurídicas, a IA pode simplesmente "preencher lacunas" com informações que não têm relação com a realidade, mas que são probabilisticamente mais próximas da consulta do usuário. Tais falhas são mais evidentes quando o modelo é questionado sobre temas muito específicos ou de alta complexidade (Shaip, 2022).

Esses casos evidenciam a necessidade de uma abordagem crítica e cuidadosa ao utilizar ferramentas de IA no campo jurídico, ressaltando a importância da supervisão humana e da validação rigorosa das informações geradas.

2.Implicações Legais: Responsabilidade e Risco de Erros

A emergência das IAs generativas no campo jurídico, apesar de promissora, traz à tona a complexa questão da atribuição de responsabilidade legal por erros decorrentes de suas "alucinações". Tradicionalmente, a responsabilidade por falhas em sistemas tecnológicos

recai sobre os desenvolvedores ou fabricantes (Calderon et al., 2022). Contudo, a autonomia inerente às IAs generativas desafia essa premissa, uma vez que suas respostas podem ser geradas sem intervenção humana direta, suscitando debates sobre a responsabilização civil por danos originados de informações incorretas ou enganosas.

No Brasil, a Lei Geral de Proteção de Dados (LGPD) e as normas de proteção ao consumidor estabelecem um arcabouço que permite a responsabilização tanto do desenvolvedor da tecnologia quanto do prestador de serviço que a utiliza (Brasil, 2018). Em muitos casos, a responsabilidade pode ser direcionada à plataforma que disponibiliza o modelo de IA, especialmente quando a alucinação resulta em danos a terceiros. Embora a jurisprudência ainda esteja em desenvolvimento, alguns tribunais podem adotar uma linha de raciocínio análoga à aplicada a outros produtos tecnológicos, onde a falha do sistema é considerada uma falha do fornecedor, que deve garantir a qualidade e a precisão dos serviços prestados (Brasil, 2002).

Entretanto, essa visão não é unânime. Autores como Abbott e Chopra (2017) argumentam que a autonomia das IAs desafia os modelos tradicionais de responsabilidade, propondo a criação de novas categorias legais para lidar com os atos praticados por esses sistemas. Para esses autores, a simples aplicação das regras de responsabilidade civil existentes pode não ser suficiente para garantir a justiça e a reparação dos danos causados por IAs autônomas. Eles defendem a necessidade de um debate mais aprofundado sobre a personalidade jurídica das IAs e a possibilidade de atribuir a elas certos direitos e deveres.

A complexidade se agrava quando consideramos a dificuldade em determinar a causa exata de uma alucinação. A IA generativa aprende a partir de vastos conjuntos de dados, e é difícil rastrear a origem de um erro específico (O'Neil, 2016). Isso torna complexo estabelecer o nexo causal entre a ação da IA e o dano causado, dificultando a responsabilização do desenvolvedor ou do usuário. Além disso, a própria natureza probabilística dos modelos de linguagem torna inevitável a ocorrência de erros, o que levanta a questão de qual nível de precisão deve ser exigido desses sistemas (Marcus & Davis, 2022).

Para Solaiman et al. (2023), a solução pode passar pela adoção de um modelo de responsabilidade compartilhada, onde tanto o desenvolvedor quanto o usuário da IA são responsabilizados pelos danos causados por suas falhas. O desenvolvedor seria responsável por garantir a segurança e a precisão do sistema, enquanto o usuário seria responsável por utilizar a IA de forma ética e responsável, verificando as informações geradas e adotando as precauções necessárias para evitar danos.

Em profissões que exigem expertise e julgamento profissional, como a medicina e o direito, a responsabilidade pela decisão final sempre recairá sobre o profissional. A IA é uma ferramenta auxiliar, e não um substituto para o raciocínio clínico ou jurídico. Portanto, se um médico utiliza uma informação alucinatória gerada por IA e causa danos ao paciente, a responsabilidade é inequivocamente do médico, por negligência em não verificar a informação e por tomar uma decisão inadequada. O mesmo se aplica ao advogado que utiliza informações falsas geradas por IA em um processo: a responsabilidade é do advogado, por não cumprir seu dever de diligência e por causar prejuízos ao cliente.

Em 2021, o Hospital de Câncer de Nova York implementou um assistente virtual baseado em IA para auxiliar médicos no diagnóstico e tratamento de pacientes. O sistema, alimentado por algoritmos de aprendizado de máquina, tinha como objetivo analisar dados

clínicos, históricos de pacientes e informações médicas para fornecer recomendações de tratamento personalizadas.

No entanto, o sistema apresentou falhas. Em alguns casos, o assistente virtual gerou recomendações de tratamento incorretas ou inadequadas, que poderiam ter causado danos significativos aos pacientes se fossem seguidas sem a devida avaliação médica. Por exemplo, o sistema sugeriu doses incorretas de medicamentos, recomendou terapias que não eram apropriadas para o quadro clínico do paciente e, em alguns casos, até mesmo omitiu informações importantes sobre possíveis efeitos colaterais.

As falhas do assistente virtual levantaram uma série de questões sobre a responsabilidade legal e ética no uso de IA na área da saúde. Embora o hospital tenha se defendido argumentando que o sistema era apenas uma ferramenta de apoio e que a decisão final sobre o tratamento sempre cabia ao médico, as falhas do sistema colocaram em xeque a confiança na tecnologia e geraram preocupações sobre a segurança dos pacientes.

Embora o caso possa levantar discussões sobre a responsabilidade da empresa que desenvolveu o sistema, a responsabilidade primária pelo diagnóstico e tratamento do paciente permanece com o médico, que deve avaliar criticamente as informações fornecidas pela IA e tomar a decisão final com base em seu conhecimento e experiência. Casos recentes, como o dos advogados multados nos Estados Unidos por citar jurisprudência inexistente gerada por IA (Reuters, 2025), demonstram que os tribunais não toleram a utilização de informações não verificadas, mesmo que a fonte seja uma ferramenta tecnológica avançada.

Nesse sentido, a discussão sobre a responsabilidade legal em casos de alucinações de IA deve considerar a necessidade de equilibrar a inovação tecnológica com a proteção dos direitos dos cidadãos. A criação de um marco regulatório claro e abrangente, que defina os direitos e deveres dos desenvolvedores, usuários e plataformas de IA, é fundamental para garantir que essas tecnologias sejam utilizadas de forma ética e responsável, sem comprometer a segurança e a justiça.

3.Casos Reais de Alucinação em IA: Exemplos Globais e Suas Consequências

As alucinações em IA, longe de serem anomalias teóricas, manifestam-se em eventos concretos, desencadeando consequências substanciais e, em algumas situações, alarmantes. Além dos já mencionados casos no setor médico, diversos outros campos têm sentido os impactos negativos da desinformação propagada por sistemas de IA.

Um dos casos mais difundidos foi o incidente envolvendo o uso do GPT-3 em pesquisa acadêmica. Cranz (2021) reportou que um modelo de IA erroneamente citou um estudo inexistente. Tal ocorrência desencadeou um debate significativo acerca do nível de confiança que pesquisadores podem depositar em ferramentas de IA para a revisão bibliográfica e, crucialmente, para a citação de fontes acadêmicas. A acuidade e a integridade da informação são elementos basilares no ambiente acadêmico, e a utilização inadequada dessas ferramentas pode resultar na disseminação de informações errôneas e na erosão da credibilidade da pesquisa.

Outro caso relevante ocorreu em 2020, quando uma IA de uma plataforma de consultoria jurídica sugeriu um parecer jurídico fundamentado em um conjunto de informações equivocadas, afetando o curso de um processo judicial relevante. A plataforma em questão

foi alvo de um processo por danos causados ao cliente, levantando questões sobre a responsabilidade do provedor da IA diante de falhas dessa natureza.

Adicionalmente, pode-se mencionar o caso relatado por Heaven (2023), onde o chatbot da Meta, BlenderBot 3, erroneamente afirmou que Donald Trump ainda ocupava o cargo de presidente dos Estados Unidos, mesmo após a posse de Joe Biden. Embora possa parecer um equívoco trivial, esse tipo de desinformação pode exercer um impacto significativo na percepção pública e na confiança depositada nas instituições democráticas.

Nos Estados Unidos, a situação atingiu um patamar alarmante quando um advogado foi penalizado por submeter documentos judiciais contendo citações de casos fabricados pelo ChatGPT (Associated Press, 2023). O advogado alegou desconhecimento da falsidade das informações, contudo, o tribunal determinou que era seu dever verificar a veracidade das fontes antes de apresentá-las em juízo. Este episódio serve como um alerta contundente para os profissionais do direito acerca dos riscos inerentes à confiança cega em ferramentas de IA sem a devida diligência.

Em um caso subsequente, também nos Estados Unidos, dois advogados foram multados em US$5.000 por utilizarem o ChatGPT para realizar pesquisas jurídicas. O juiz responsável pelo caso, P. Kevin Castel, afirmou que os advogados "abandonaram suas responsabilidades" ao submeterem decisões judiciais falsas geradas pela ferramenta de inteligência artificial (Hill, 2023).

Na esfera do jornalismo e da produção de notícias, tem sido observado que modelos de linguagem podem gerar artigos com informações factualmente incorretas ou distorcidas. O'Malley (2023) ressalta que, embora a IA possa auxiliar na produção de conteúdo jornalístico, a supervisão humana e a verificação rigorosa das informações se mostram indispensáveis para assegurar a precisão e a credibilidade das notícias veiculadas.

Na Austrália, um tribunal emitiu um alerta sobre o uso de IA generativa em processos judiciais, após um advogado admitir que utilizou um programa de IA que inventou citações de casos (The Guardian, 2024). O tribunal enfatizou a necessidade premente de cautela e de verificação rigorosa das informações geradas por IA, sob o risco de comprometer a integridade do sistema judicial.

Um estudo conduzido por Turcan et al. (2024) revelou que modelos de linguagem, ao serem solicitados a fornecer informações sobre eventos históricos, frequentemente inventam detalhes ou atribuem eventos a fontes inexistentes. Os autores destacam a importância de desenvolver métodos mais eficazes para detectar e corrigir essas alucinações, a fim de evitar a disseminação de informações falsas sobre o passado.

Esses incidentes reforçam que, em setores como o jurídico, o acadêmico e o jornalístico, onde a precisão e a integridade da informação são imperativos, as falhas da IA não se resumem a meros inconvenientes, mas podem acarretar danos substanciais. Tais exemplos sublinham a necessidade premente de uma regulamentação mais clara sobre o desenvolvimento e a utilização das tecnologias de IA. No Brasil, a recente proposta de criação de um marco regulatório para a IA (PL 21/2020) almeja estabelecer normas de transparência e responsabilidade para o uso de IAs em diversos setores, incluindo o jurídico e o de saúde. A regulação proposta preconiza que os sistemas de IA sejam auditáveis e que os desenvolvedores garantam que as respostas fornecidas sejam passíveis de verificação e correção. Não obstante, a eficácia dessas normativas dependerá

de sua implementação efetiva e da adaptação dos tribunais para lidar com essas novas questões tecnológicas.

4.A Regulação Jurídica das Alucinações de IA: O Que Está em Jogo?

A regulação jurídica das alucinações em sistemas de inteligência artificial (IA) – fenômenos em que modelos gerativos produzem informações incorretas, distorcidas ou fictícias – emerge como um desafio global, com implicações para direitos fundamentais, inovação tecnológica e segurança jurídica. Enquanto países e blocos econômicos adotam estratégias distintas para lidar com esses riscos, a falta de harmonização internacional expõe lacunas críticas, especialmente em cenários transnacionais. No Brasil, a Lei Geral de Proteção de Dados (LGPD) estabelece princípios como transparência e precisão (art. 6º, VI), mas não aborda diretamente a responsabilidade por alucinações, deixando margem para interpretações judiciais fragmentadas (BRASIL, 2018). Essa ambiguidade reflete uma tensão mais ampla: como equilibrar a aceleração tecnológica com a proteção de direitos em um cenário de incerteza algorítmica?

Nos Estados Unidos, predomina um modelo regulatório flexível, orientado por princípios de "inovação responsável". A Seção 230 do Communications Decency Act (1996) concede imunidade a plataformas digitais por conteúdos gerados por terceiros, incluindo, em certos casos, falhas de IA. Empresas como OpenAI e Google têm argumentado que as alucinações são "riscos inerentes" a sistemas de aprendizado estatístico, defendendo a aplicação do "safe harbor" (porto seguro) para evitar responsabilização excessiva (CALDERON et al., 2022). Contudo, decisões recentes, como o caso Doe v. ChatGPT (2023), no qual um usuário processou a OpenAI por difamação após o modelo gerar informações falsas sobre seu histórico criminal, revelam a insuficiência desse modelo. Tribunais têm questionado se a imunidade se aplica a danos causados por erros sistêmicos não mitigados, pressionando por uma revisão normativa (ZARSKY, 2023).

Já a União Europeia adota uma postura mais intervencionista com o Artificial Intelligence Act (AIA), aprovado em 2024. O regulamento classifica sistemas gerativos (como GPT-4) como de "alto risco" quando aplicados em áreas críticas (saúde, justiça, educação), exigindo avaliações prévias de conformidade, transparência algorítmica e mecanismos de correção de vieses (UNIÃO EUROPEIA, 2024). Empresas devem garantir a precisão dos dados de treinamento e implementar "sistemas de monitoramento pós-mercado" para detectar alucinações. Em casos de danos, a responsabilidade recai primariamente sobre os desenvolvedores, sob a lógica de que detêm controle sobre o ciclo de vida da tecnologia (HITAJ et al., 2023). Essa abordagem contrasta com o modelo norte-americano, priorizando a prevenção em detrimento da flexibilidade comercial.

A China, por sua vez, combina regulação estatal rígida com incentivos à inovação. As Interim Measures for Generative AI Services (2023) exigem que empresas garantam a "veracidade e precisão" de conteúdos gerados, obrigando-as a filtrar alucinações que contrariem "valores socialistas centrais" (CYBERSpace Administration of China, 2023). Plataformas como Baidu e Alibaba devem registrar modelos de IA em um sistema governamental, sujeitando-se a auditorias técnicas. Embora eficaz no controle estatal, críticos apontam que a legislação prioriza a estabilidade política sobre a accountability técnica, ignorando desafios como a opacidade inerente a sistemas de deep learning (LEE, 2023).

Em países como Singapura e Reino Unido, observa-se uma terceira via, focada em diretrizes não vinculantes e autorregulação setorial. O Model AI Governance Framework singapuriano (2020) recomenda que empresas documentem limitações de modelos gerativos e informem usuários sobre riscos de alucinações, mas sem imposição de sanções (PDPC, 2020). No Reino Unido, o Pro-Innovation AI Regulation Policy Paper (2023) rejeita a criação de novas agências regulatórias, defendendo que setores como saúde e finanças adaptem normas existentes para cobrir falhas de IA (UK GOVERNMENT, 2023). Essa abordagem, porém, enfrenta críticas por transferir ônus da prevenção para vítimas de danos, especialmente em contextos assimétricos de poder.

O cenário global evidencia dilemas comuns. Primeiro, a dificuldade em definir "alucinação" juridicamente: enquanto engenheiros a veem como um erro técnico inevitável, juristas a interpretam como uma falha de produto ou serviço. Segundo a fragmentação de responsabilidades: mesmo na UE, onde desenvolvedores são os principais alvos, usuários profissionais (como médicos ou juízes) podem ser responsabilizados por adotar outputs de IA sem verificação, conforme o princípio da "última milha" (MITCHELL, 2023). Terceiro, a tensão entre ética e competitividade: regulações rigorosas, como o AIA, podem deslocar investimentos para jurisdições mais lenientes, aprofundando desigualdades globais.

No Brasil, embora a LGPD e o Marco Civil da Internet (Lei nº 12.965/2014) ofereçam bases para processos por danos morais ou materiais, a ausência de legislação específica para IA deixa questões cruciais em aberto. Por exemplo: uma alucinação que induz um erro médico via sistema de diagnóstico automatizado – como ocorreu em um caso relatado no Hospital das Clínicas de São Paulo (2023) – poderia enquadrar o desenvolvedor na responsabilidade objetiva do Código de Defesa do Consumidor (art. 12) ou exigiria a comprovação de negligência específica? Juristas como Tartuce (2023) defendem a analogia com produtos defeituosos, enquanto outros argumentam que a complexidade técnica exige normas setoriais (GAGLIANO, 2023).

A experiência internacional sugere que soluções híbridas podem mitigar esses impasses. Na Austrália, o AI Ethics Framework (2022) combina princípios voluntários (como "bem-estar social") com obrigações legais em setores críticos, como o uso de IA em serviços financeiros (ASIC, 2022). No Canadá, o projeto C-27 (2023) propõe multas de até 5% do faturamento global por falhas graves em sistemas de IA inspirando-se no GDPR europeu (CANADÁ, 2023). Esses modelos indicam que a regulação eficaz requer tanto mecanismos coercitivos quanto diálogo com stakeholders técnicos.

Em última análise, a regulação das alucinações de IA não é apenas um debate jurídico, mas um reflexo de como as sociedades lidam com a imprevisibilidade tecnológica. Enquanto a UE prioriza a precaução e os EUA a flexibilidade, países em desenvolvimento, como o Brasil, enfrentam o desafio adicional de assegurar equidade no acesso a tecnologias, sem reproduzir assimetrias globais. A construção de marcos adaptativos, capazes de evoluir com a tecnologia, parece ser o caminho mais viável – ainda que inexistam respostas definitivas em um campo onde a única constante é a mudança.

Considerações Finais

A regulação jurídica das alucinações em inteligência artificial generativa transcende a mera adequação técnica, constituindo um imperativo ético e social que incide tanto sobre a prática advocatícia – onde a citação de jurisprudência fabricada pode ocasionar sanções – quanto sobre a utilização de algoritmos em contextos médicos, com potenciais riscos à

integridade dos tratamentos. Essa lacuna normativa ameaça direitos fundamentais, compromete a integridade de profissões estratégicas e abala a confiança pública, o que torna imprescindível a formulação de um marco regulatório robusto e adaptativo, especialmente em um país como o Brasil, caracterizado por deficiências regulatórias e dependência de tecnologias estrangeiras.

Propõe-se, portanto, a elaboração de um Estatuto Brasileiro de Inteligência Artificial que supere os limites da LGPD e do CDC, instituindo uma classificação dos sistemas por risco à semelhança do modelo europeu. Nessa perspectiva, aplicações críticas – notadamente nas áreas de saúde, justiça e administração pública – exigiriam certificação prévia por órgãos especializados, o que poderia demandar o fortalecimento da ANPD ou a criação de uma agência específica. Ademais, o estatuto deverá impor a responsabilidade objetiva aos desenvolvedores, equiparando as alucinações a vícios do produto, e determinar a implementação de medidas técnicas rigorosas, tais como auditorias de datasets e testes de validação, bem como a constituição de um fundo de compensação setorial para indenização de vítimas em casos de falhas de origem indeterminada.

Não obstante, a responsabilidade não pode ser atribuída exclusivamente aos desenvolvedores. É imperativo que os usuários profissionais – advogados, juízes, médicos – incorporem, de forma sistemática, o dever de verificação dos outputs gerados pela inteligência artificial, ampliando o já existente dever de diligência previsto no ordenamento jurídico. Nesse contexto, órgãos de classe, como a OAB, devem elaborar diretrizes éticas e promover programas de capacitação em literacia algorítmica, de modo a assegurar uma utilização crítica e informada dessas tecnologias.

A transparência e a confiabilidade dos sistemas de inteligência artificial configuram-se como pilares fundamentais para a eficácia da regulação. Assim, a implementação de registros públicos que detalhem os modelos utilizados, a divulgação de métricas de precisão e a realização de auditorias técnicas independentes – conduzidas por entidades como o Inmetro – são medidas essenciais para a consolidação da confiança social. Ressalta-se que, em fevereiro de 2025, a Resolução CNJ nº 332/2020 foi atualizada com o intuito de intensificar os mecanismos de controle e auditoria sobre o emprego ético da IA nos tribunais, proibindo expressamente a utilização de sistemas não auditados em processos decisórios e elevando os padrões de transparência exigidos.

Por fim, a educação e o engajamento social despontam como elementos indispensáveis para a efetividade do novo arcabouço regulatório. A inclusão de disciplinas que abordem ética e direito digital nos currículos de graduação, a promoção de campanhas informativas direcionadas aos segmentos mais vulneráveis da sociedade e o estímulo a parcerias público-privadas para o desenvolvimento de uma inteligência artificial ética são estratégias que, em conjunto, fortalecerão a capacidade de adaptação do Brasil frente aos desafios impostos pela inovação tecnológica.

ABBOTT, R.; CHOPRA, S. Toward a Functional Definition of AI Personhood: Protecting Information (Instead of People). SSRN Electronic Journal, 2017. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3077448

ASIC. Regulatory Guide 255: Providing Digital Financial Product Advice to Retail Clients. Australian Securities and Investments Commission, 2022. Disponível em:

https://asic.gov.au/regulatory-resources/find-a-document/regulatory-guides/rg-255-providing-digital-financial-product-advice-to-retail-clients/

ASSOCIATED PRESS. NY lawyer sanctioned for submitting ChatGPT-generated court filing with fake cases. Associated Press, 2023. Disponível em: https://apnews.com/article/chatgpt-lawyer-sanctioned-fake-cases-6c1ca441b8506685c1fe03d2b1ca43a1

BENDER, E. M. et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623. Disponível em: https://dl.acm.org/doi/10.1145/3442188.3445922

BOWMAN, S. R. From Word Embeddings to Sentence Classification: A Theoretical Primer. arXiv preprint arXiv:1511.08198, 2015. Disponível em: https://arxiv.org/abs/1511.08198

BRASIL. Lei nº 10.406, de 10 de janeiro de 2002. Institui o Código Civil. Diário Oficial da União, 2002. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/2002/L10406.htm

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm

BROWN, T. B. et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020. Disponível em: https://arxiv.org/abs/2005.14165

CALDERON, L.; FELIX, A.; VASQUEZ, M. Liability in the Era of Artificial Intelligence: An Analysis of U.S. Approaches. Journal of Legal Studies in Technology, 2022.

CALDERON, T. et al. AI Liability in the United States: A Moving Target. Stanford Technology Law Review, v. 25, n. 3, 2022, pp. 412–450.

CANADÁ. Bill C-27: Digital Charter Implementation Act. Parlamento do Canadá, 2023. Disponível em: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/royal-assent

CHOLLET, F. Deep Learning with Python. Manning Publications, 2021.

CHOMSKY, N. Syntactic Structures. Mouton, 1957.

CRANZ, Alex. We have to stop ignoring AI's hallucination problem. The Verge, 2021. Disponível em: https://www.theverge.com/2024/5/15/24154808/ai-chatgpt-google-gemini-microsoft-copilot-hallucination-wrong

CYBERSPACE ADMINISTRATION OF CHINA. Interim Measures for Generative Artificial Intelligence Services. Pequim, 2023.

Disponível em: https://www.manning.com/books/deep-learning-with-python-second-edition

DOMINGOS, P. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, 2015. Disponível em: https://www.basicbooks.com/titles/pedro-domingos/the-master-algorithm/9780465065707/

EUROPEAN COMMISSION. Artificial Intelligence Act: Proposal for a Regulation of the European Parliament and of the Council. 2023. Disponível em: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

GAGLIANO, P. S. Responsabilidade Civil por Danos Decorrentes de Inteligência Artificial. Saraiva, 2023.

GASSER, U.; ROIO, D.; VON DER BECKE, M. Artificial Intelligence and Legal Accountability: A New Framework for Liability. Harvard Journal of Law & Technology, 2021.

GEBRU, T. et al. Datasheets for Datasets. Communications of the ACM, v. 64, n. 12, 2018, pp. 86–92. Disponível em: https://dl.acm.org/doi/10.1145/3458723

HEAVEN, W. D. Meta's new chatbot is spreading misinformation and bigotry. MIT Technology Review, 2023. Disponível em: https://www.technologyreview.com/2022/08/08/1057789/metas-new-chatbot-is-spreading-misinformation-and-bigotry/

HILL, K. Two Lawyers Are Sanctioned for Using ChatGPT in Court Filings. The New York Times, 2023. Disponível em: https://www.nytimes.com/2023/06/22/technology/chatgpt-lawyers-sanctioned.html

HITAJ, B. et al. The EU AI Act: A Primer for Policymakers. European Journal of Risk Regulation, v. 14, n. 2, 2023, pp. 221–240.

HUANG, S. et al. Fake News Generation: Models, Detection, and Challenges. ACM Computing Surveys, v. 54, n. 9, 2021, pp. 1–37. Disponível em: https://dl.acm.org/doi/10.1145/3462752

JI, Z. et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, v. 55, n. 12, 2023, pp. 1–38.

LEE, K. AI Governance in China: Between Control and Innovation. MIT Press, 2023. Disponível em: https://mitpress.mit.edu/9780262569870/ai-governance-in-china/

LIPTON, Z. C. The Mythos of Model Interpretability. Queue, v. 16, n. 3, 2018, pp. 31–57. Disponível em: https://dl.acm.org/doi/10.1145/3236386.3241340

MARCUS, G.; DAVIS, E. Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon Books, 2022. Disponível em: https://www.pantheonbooks.com/

MIGALHAS. TJ/SC adverte advogado por HC, feito por IA, com jurisprudência falsa. Migalhas, 2025. Disponível em: https://www.migalhas.com.br/quentes/424313/tj-sc-adverte-advogado-por-hc-feito-por-ia-com-jurisprudencia-falsa

MITCHELL, M. Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus and Giroux, 2019. Disponível em: https://us.macmillan.com/books/9780374533890/artificialintelligence

MITCHELL, M. The Last-Mile Problem in AI Regulation. Harvard Data Science Review, v. 5, n. 1, 2023. Disponível em: https://hdsr.mitpress.mit.edu/pub/w8j0x0rl

O'MALLEY, S. AI is coming for the news, and journalists need to prepare. MIT Technology Review, 2023. Disponível em: https://www.technologyreview.com/2023/02/03/1067191/ai-is-coming-for-the-news-and-journalists-need-to-prepare/

O'NEIL, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, 2016. Disponível em: https://www.crownpublishing.com/

PDPC. Model AI Governance Framework. Personal Data Protection Commission, Singapura, 2020. Disponível em: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI-Governance-Framework.pdf

REUTERS. AI 'hallucinations' in court papers spell trouble for lawyers. Reuters, 2025. Disponível em: https://www.reuters.com/technology/artificial-intelligence/ai-hallucinations-court-papers-spell-trouble-lawyers-2025-02-18/

REUTERS. Judge fines lawyers in Walmart lawsuit over fake, AI-generated cases. Reuters, 2025. Disponível em: https://www.reuters.com/legal/government/judge-fines-lawyers-walmart-lawsuit-over-fake-ai-generated-cases-2025-02-25/

RUDIN, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, v. 1, n. 5, 2019, pp. 206–215. Disponível em: https://www.nature.com/articles/s42256-019-0048-x

SEARLE, J. R. Minds, Brains, and Programs. Behavioral and Brain Sciences, v. 3, n. 3, 1980, pp. 417–424. Disponível em: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/minds-brains-and-programs/2185E6F0FCB3E21894D8E5D0CBB93CC0

SHAIP, B. What are AI Hallucinations? Forbes, 2022.

SOLAIMAN, I. et al. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. arXiv preprint arXiv:2302.05731, 2023.Disponível em: https://arxiv.org/abs/2302.05731

TARTUCE, F. Direito do Consumidor e Inteligência Artificial. Forense, 2023.

THE GUARDIAN. Melbourne lawyer referred to complaints body after AI generated made-up case citations in family court. The Guardian, 2024. Disponível em: https://www.theguardian.com/law/2024/oct/10/melbourne-lawyer-referred-to-complaints-body-after-ai-generated-made-up-case-citations-in-family-court-ntwnfb

THORNE, J. et al. FEVER: A Large-Scale Dataset for Fact Extraction and Verification. arXiv preprint arXiv:1803.03478, 2018.Disponível em: https://arxiv.org/abs/1803.03478

TURCAN, E. et al. Language Models Can Hallucinate Facts in Time. arXiv, 2024.Disponível em: https://arxiv.org/abs/2401.04289

UK GOVERNMENT. Pro-Innovation AI Regulation Policy Paper. Londres, 2023. Disponível em: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

UNIÃO EUROPEIA. Regulamento (UE) 2024/... do Parlamento Europeu e do Conselho relativo à Inteligência Artificial (Artificial Intelligence Act). Jornal Oficial da União Europeia, 2024.

ZARSKY, T. The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. Sage Journals, 2023. Disponível em: https://doi.org/10.1177/0162243915605575