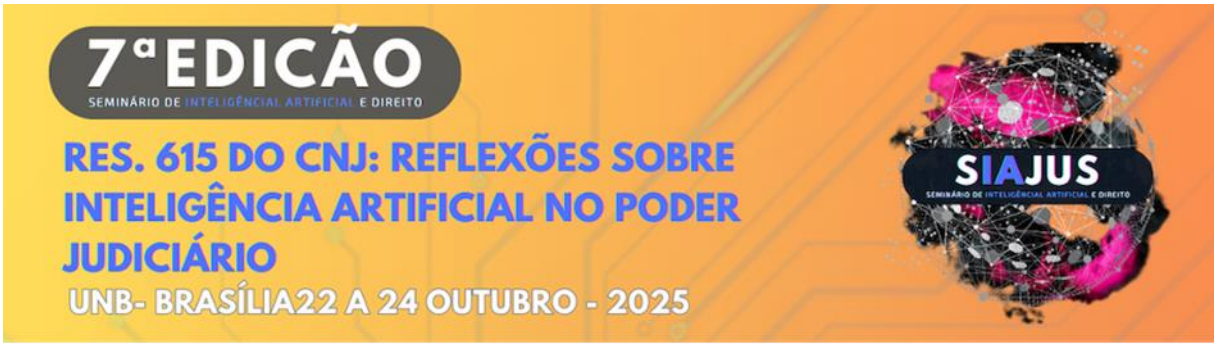


IAgen SEM ALUCINAÇÃO: SEGURANÇA NO PROCESSO JUDICIAL.

S. Tavares-Pereira[1]

Palavras-chave: Inteligência artificial generativa; confabulação; entropia semântica; entropia estatística; metacognição.

INTRODUÇÃO

Há um momento operativo decisivo em que os algoritmos de IA generativa (IAgen) entram numa "zona de perigo de alucinações". Esse é o achado central de pesquisadores de Oxford (FARQUHAR, 2024) conforme o relatório publicado na revista *Nature* e objeto desta *review*. As alucinações corroem a confiabilidade da IA e evitá-las interessa muito à comunidade jurídica do Brasil porque o Conselho Nacional de Justiça (CNJ, 2025) autorizou o uso amplo dessa ferramenta no suporte à decisão e sinalizou, assim, uma inflexão histórica na tradicional resistência do sistema jurídico às inovações disruptivas.

Estar ciente da existência dessas ferramentas, ainda mal-entendidas e em franca evolução, e dispor-se a usá-las, apesar dos riscos, são as ideias guias desta revisão que alerta para os limites de uso da IA na decisão judicial. Roteiro: conceitos operacionais, visão geral da pesquisa, questões excluídas da pesquisa, mas relevantes para o Direito e considerações finais.

1. CONCEITOS OPERACIONAIS.

Para os fins deste texto, considera-se: **(i)** confabulação (80% das alucinações): o comportamento dos algoritmos que, diante de incerteza estatística, forjam uma resposta arbitrária e errada, sensível a detalhes irrelevantes; o fenômeno é análogo à confabulação psicológica; **(ii)** Entropia: a medida da indiferenciação ou desorganização de elementos em um sistema, sendo seu oposto a organização, que exige esforço e energia (neguentropia); **(iii)** Entropia estatística/probabilística: quando todas as alternativas disponíveis ao algoritmo são igualmente prováveis, gerando perplexidade e risco de confabulação; **(iv)** Entropia semântica:

[1] Mestre em Ciência Jurídica/Univali-SC e doutorando pela Atitus/FDV. E-mail: stavarespereira@gmail.com. Currículo Lattes: <http://lattes.cnpq.br/7899125210261357>.
.

obtida com o refinamento linguístico da entropia estatística e, consequentemente, decidir pela sequência ou interrupção do diálogo (evitar a alucinação/confabulação); **(v)** *Semantic clustering:* operação de classificação de elementos probabilisticamente similares com técnica como a da implicação bidirecional que comprova a equivalência semântica ou não; **(vi)** metacognição ou intervenção metacognitiva: neste contexto, designa interferências externas diretas na estrutura cognitiva dos algoritmos aprendizes, permitindo ajustes operacionais induzidos de fora, por analogia com o conceito psicológico de "pensar sobre o próprio pensar", embora sem consciência ou autocrítica, como ocorre em sistemas psíquicos; **(vii)** vetorização: técnica matemática que converte dados linguísticos não estruturados em representações vetoriais de alta dimensionalidade, permitindo contextualizar significados com base na proximidade entre palavras ou sentenças em espaços semânticos. Implementada em abordagens como Word2Vec (vetorização por palavras) e Sent2Vec (por sentenças), inspiradas na filosofia da linguagem, que conecta significado ao uso contextual e **(viii)** implicação bidirecional: relação entre termos ou sentenças semanticamente equivalentes, intercambiáveis nos contextos. Permite formar *clusters* semanticamente entrópicos e há ferramenta para isso nativa nos LLMs.

2. A PESQUISA

A linguagem natural já é manipulada com precisão por algoritmos aprendizes capazes de extrair padrões cognitivos e comportamentais dos dados. A identificação de recorrências (padrões) em dados, com fins instrumentais, desafia o antigo princípio de que de fatos não se deduzem normas. Destilando os dados e captando recursividades algoritmizáveis, a IA revela estruturas de ação, interpreta-as como regras implícitas e as codifica para uso próprio.

Contudo, esses sistemas ainda estão em fase de amadurecimento e apresentam falhas, como as alucinações (GOSMAR e DAHL, 2025; GERCINA, 2025), fenômeno que pesquisadores de Oxford explicaram, com ênfase para um subconjunto (80% delas) que denominaram confabulações. A pesquisa (FAQUHAR, 2024) identifica as condições internas de processamento que levam ao problema e sugere técnica para resolvê-lo: *clustering* semântico com base em mecanismo de implicação bidirecional, nativo nos LLMs.

Os desvios alucinógenos não decorrem de uma suposta "autonomia" dos algoritmos, mas de limitações técnicas na mecânica operativa dos modelos. A proposta dos cientistas é uma

espécie de intervenção metacognitiva: adestrar os modelos para que "pensem" estritamente, como máquinas inteligentes, capazes de prever e evitar comportamentos que tornem seus resultados inúteis ou perigosos. Sugere-se o aperfeiçoamento heurístico do algoritmo.

Além da otimização heurística do algoritmo (nova função introduzida no código-fonte), a pesquisa alerta para causas externas, indutoras de erro, de que não se ocuparam, e que exigem atenção dos juristas visando modelagens jurídicas e constitucionais mais seguras.

3. MELHORANDO A HEURÍSTICA DO ALGORITMO APRENDIZ.

A pesquisa dos cientistas de Oxford, também divulgada por TIMMER (2024), propõe uma técnica para detectar a ocorrência da situação operacional (estado interno das variáveis) em que a IAgen pode confabular. A hipótese central, confirmada ao final do experimento, foi: a confabulação é precedida de um momento de incerteza sobre fatos ou formulações. Desamparado pela estatística e sem uma opção claramente "mais provável" para continuar a interação determinística em curso, o algoritmo faz escolhas e chega a construções arbitrárias (confabula). A lógica de montagem de uma frase permite explicar a técnica: se a análise probabilística dos "termos candidatos para serem o próximo" evidencia um mais provável[2], o algoritmo o usa e prossegue; se não, o cenário é de indiferenciação entrópica - quando há muitos termos com idênticas probabilidades - e o risco de confabulação aumenta. Os pesquisadores propõem, então, uma classificação adicional dos termos, baseada em implicação bidirecional (agrupar os termos por significado). Se essa *clusterização* gera múltiplos conjuntos, o risco de confabulação é alto e o algoritmo deve interromper o diálogo. Quebra-se, assim, o determinismo maquínico de produção de um resultado. Se é gerado apenas um conjunto semântico, qualquer termo pode ser usado com segurança. Essa abordagem identifica o ponto de passagem "da incerteza para a invenção" e oferece uma ferramenta concreta para prevenir riscos e produzir uma IA mais confiável (REGULAMENTO (UE), 2024). Trata-se de um aperfeiçoamento heurístico (TAVARES-PEREIRA, 2021, p. 281) do algoritmo, de valor inestimável na atual fase de receptividade à IAgen. A pesquisa expande a possibilidade de uso confiável e ético da IAgen em contextos processuais.

[2] A probabilidade de que seja ele o termo adequado é, percentualmente, bem superior a outros.

4. EXCLUSÕES DA PESQUISA, MAS NÃO DAS PREOCUPAÇÕES.

A pesquisa concentrou-se nas causas técnicas da confabulação da IAgen, sanáveis pela melhoria do código algorítmico. E excluiu, intencional e metodologicamente, fatores responsáveis por cerca de 20% das respostas falsas geradas pelos diferentes modelos, os quais apontam para interferências externas que podem comprometer o aprendizado, como: seleção inadequada das bases de dados de treinamento[3], fraude no processo de treinamento[4] e limitação técnica[5] do próprio mecanismo de aprendizado (algoritmo). Tais interferências — que poderiam ser denominadas de metacognição e/ou tecno-indução — podem enviesar as estruturas operativas algoritmizadas. Bases de dados idôneas e supervisão de treinamento adequada e ética, em conjunto, permitem gerar uma IA confiável. A qualidade dos algoritmos aprendizes utilizados completam a "tríade do mais confiável desempenho". Ela permite incorporar a IAgen ao sistema jurídico com a garantia de um aprendizado autônomo e eficaz. A tríade **(i)** reforça a necessidade de vigilância, discernimento e institucionalização das escolhas para o preparo dos modelos e **(ii)** demonstra a falácia da ideia generalizada de que a IA é uma caixa preta. Uma parte substancial da caixa é algoritmo clássico, cuja heurística é continuamente aperfeiçoada.

---

[3] Enviesamento e base de treinamento: a base condiciona diretamente as estruturas operativas/cognitivas autoconstruídas pelo algoritmo aprendiz (bases de conhecimento/funções de transformação – a caixa-preta). Assim, precisa ser institucionalmente constituída, sem exclusão de fatos/procedimentos reais e sem inclusão de falsidades. Mecanismos democráticos devem orientar a formação, para que espelhe a pluralidade e a diversidade, por exemplo, e exclua vieses. Emular o agir dos juízes humanos e seu "livre convencimento motivado" exige modelagens que levem a IA até onde deve/pode ir no processo judicial (suporte ao juiz/CNJ 615). A seleção da base deve ser institucionalizada para, sendo idônea, garantir representatividade e legitimidade.

[4] A fraude no processo de treinamento, ou tecno-indução, não se confunde com a manipulação da base de treinamento. De uma base idônea podem surgir estruturas enviesadas pela interferência da supervisão. Exemplo: manipular parâmetros das redes neurais (pesos), em geral pelo ajuste do "bias", e forçar a convergência do algoritmo para um resultado. A técnica legítima pode ser usada para sub valorar/sobre valorar variáveis e enviesar o sistema. Juízes aplicam diferentes ponderações/interpretações aos mesmos fatos/normas o que gera variância nas decisões. Isso é democrático e constitucional e, portanto, um valor. A externalização do mecanismo (tecno-indução) é espúria e pode transformar o processo numa fábrica de respostas únicas e/ou enviesadas.

[5] A produção de esquemas de generalização robustos é o núcleo do *machine learning e*, também, uma de suas maiores fragilidades. Sistemas psíquicos equilibram abstração com percepção e atenção. Algoritmos operam mecanicamente. Classificam elementos com base em padrões, sem sensibilidade a nuances contextuais. Perístases (GUNTHER, 2004, p. 25) não existem para o código em operação. Gerar funções de generalização, sem excesso de parâmetros (overfitting) ou escassez (underfitting), é um desafio. Então, em decisões, o uso deve ser admitido após o juiz humano ter uniformizado os critérios, como em demandas repetitivas. O protagonism humano, em relação ao novo, tem de ser preservado no processo.

CONSIDERAÇÕES FINAIS

As confabulações (80% das alucinações) da IAGen são causadas por uma situação entrópica probabilística. Uma melhoria heurística do algoritmo aprendiz permite confirmar se a entropia é "também semântica", caso em que a operação é continuada. Se não, aborta-se o processo para evitar a confabulação. A manobra evidencia que a IA não é apenas uma imensa caixa-preta. Ela é híbrida, composta de um algoritmo clássico e de uma base de conhecimento (funções de transformação codificadas pelo algoritmo no treinamento) que, esta sim, é uma caixa-preta.

Uma tríade de fatores, de incidência simultânea e harmônica, permite gerar modelos de IAgen confiáveis: **(i)** seleção do melhor algoritmo aprendiz para a tarefa e a otimização heurística contínua dele; **(ii)** institucionalização de procedimentos de composição de bases idôneas de treinamento (plural e diversamente democráticas) e **(iii)** institucionalização e uso efetivo de mecanismos de controle/auditagem da supervisão do treinamento. O aprofundamento dessas noções e o tratamento das preocupações daí advindas, deverão ser objetos do esforço conjunto de juristas e tecnólogos.

REFERÊNCIAS BIBLIOGRÁFICAS

BRASIL. *Resolução 615/CNJ*, de 13 de março de 2025. Disponível em: <https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf>. Acesso em: 21 ago. 2025.

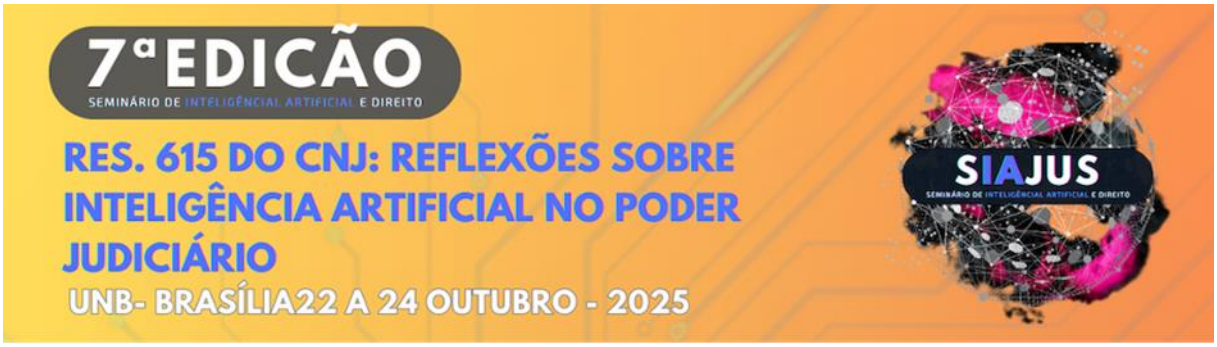
GERCINA, Cristiane. *Justiça do Trabalho condena trabalhadora após advogada usar IA e inventar decisões judiciais*. Disponível em: <https:ww1.folha.uol.com.brercado0259ustica-do-trabalho-condena-trabalhadora-após-advogada-usar-ia-e-inventar-decisoes-judiciais.shtml¿pwgt=l94mlh4jm7li843qaqws7vb9wc1ywryvjlryikgycmtzps76&utm_source=whatsapp&utm_medium=social&utm_campaign=compwagift>. Acesso em: 20 set. 2025.

GOSMAR, Diego; DAHL, Deborah A. *Hallucination mitigation using agentic AI natural language-based frameworks*. Disponível em: <https:rxiv.orgdf501.13946>. Acesso em: 10 ago. 2025.

FARQUHAR, S., KOSSEN, J., KUHN, L. *et al. Detecting hallucinations in large language models using semantic entropy*. **Nature 630**, 625–630 (2024). Disponível em: <https://doi.org/10.1038/s41586-024-07421-0>. Acesso em: 10 ago. 2025.

GÜNTHER, Klaus. **Teoria da argumentação no direito e na moral: justificação e aplicação.** São Paulo: Landy Editora, 2004.

REGULAMENTO (UE) 2024/1689 DO PARLAMENTO EUROPEU E DO CONSELHO. Disponível em: <https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX:32024R1689#document1>. Acesso em: 24 jul. 2024.



TAVARES-PEREIRA, S. **Machine learning nas decisões.** O uso jurídico dos algoritmos aprendizes. Florianópolis: Artesam. 2021. 796p.

TIMMER, John. *Researchers describe how to tell if chatGPT is confabulating*. Disponível em: <https://arstechnica.com/ai/2024/06/researchers-describe-how-to-tell-if-chatgpt-is-confabulating/>. Acesso em: 5 jul. 2024.