Shadow AI no Direito: o risco do uso da IA nas sombras

Tenho abordado [1] [2] os riscos que envolvem o uso indiscriminado e não supervisionado de inteligência artificial (IA) no Direito. O fenômeno, conhecido como *Shadow AI*, descreve o uso de ferramentas de IA sem o conhecimento, controle ou autorização formal das instituições [3].

Magnific

Na prática, trata-se de magistrados, servidores ou advogados que recorrem a plataformas de IA como assistentes para tomada de decisão, elaboração de petições ou decisões, sem qualquer rastreabilidade institucional ou cumprimento de normas de governança. A facilidade de acesso e a crescente eficiência desses sistemas contribuem para a rápida disseminação de seu uso informal, para o *workslop*,[4] criando um ambiente propício para a desregulação silenciosa e a erosão de práticas jurídicas tradicionais.

Vários fatores impulsionam o *Shadow AI* no Direito. A pressão por prazos e resultados frequentemente leva profissionais a buscar qualquer ferramenta que agilize o trabalho. A ausência de políticas claras ou soluções oficiais de IA, somado a alta usabilidade das aplicações de IA generativa, incentiva advogados e equipes jurídicas a recorrerem a plataformas gratuitas (ou pagas) disponíveis, por vezes usando contas pessoais para trabalhar com dados sensíveis. Além disso, profissionais mais jovens ou familiarizados com tecnologia tendem a adotar essas ferramentas espontaneamente, mesmo sem plena consciência dos riscos, dada a familiaridade com os *chatbots* em seu dia a dia.

O uso de IA nas sombras não é um risco imaginado. Conforme levantamento realizado pelo CNJ em parceria com o PNUD, mais de 45% dos tribunais e conselhos brasileiros declararam, em 2024, utilizar ferramentas de IA generativa em suas operações, enquanto outros 81% entre os que ainda não usam manifestaram interesse em adotá-las [5].

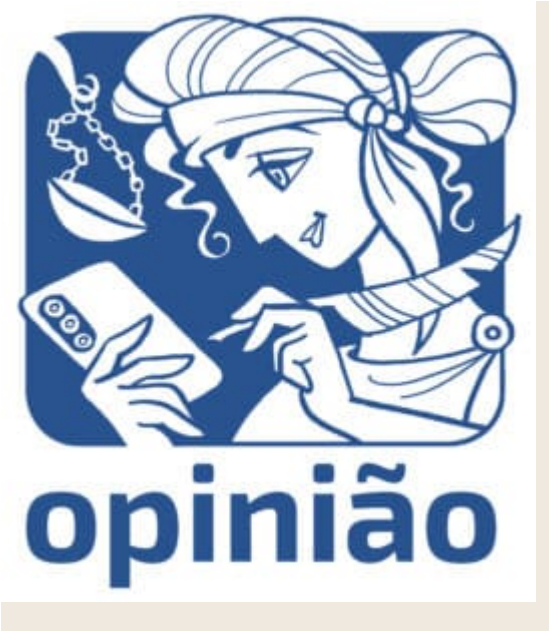
Contudo, em 57,6% dos casos em que essas ferramentas são usadas, o acesso é feito por meio de contas pessoais, sem qualquer centralização institucional. Além disso, apenas 5,9% dos órgãos do Judiciário declararam possuir diretrizes bem definidas e abrangentes para o uso dessas tecnologias. O dado revela um panorama de desgovernança, em que sistemas são utilizados sem os devidos controles e fora do âmbito normativo de segurança e responsabilidade institucional.

Essa situação se mostra particularmente grave diante dos riscos envolvidos. Ferramentas de IA generativa são sistemas tendencialmente opacos e não determinísticos, cuja base de treinamento não é transparente e cujas “saídas” podem refletir vieses, erros factuais e inconsistências.

Seu uso indiscriminado pode comprometer a imparcialidade de decisões, propagar estereótipos ou, pior, sem literacia adequada, levar a erros graves em pareceres, petições e decisões judiciais. Há ainda os riscos de vazamento de dados pessoais e sigilosos, especialmente quando usuários copiam e colam trechos de processos em plataformas abertas, muitas vezes hospedadas fora do país, sob regimes legais diversos.

A responsabilidade civil, administrativa e mesmo penal por esse uso irregular ainda é uma zona cinzenta, o que contribui para um cenário de insegurança jurídica e institucional. Incontáveis casos já foram reportados em que sistemas baseados em IA emitiram conclusões jurídicas erradas ou forneceram fundamentos falsos para decisões judiciais automatizadas (alucinações), causando constrangimento e prejuízos processuais às partes envolvidas.

Spacca



Outros riscos incluem o uso de IA para automação de tarefas críticas sem supervisão humana adequada, o que pode gerar dependência tecnológica, perda de habilidades (*deskilling*) e dificultar a apuração de responsabilidades. A ausência de padronização entre os órgãos agrava o problema, criando uma assimetria de práticas e dificultando a auditoria dos sistemas.

Não se trata de demonizar a tecnologia. A IA pode ser aliada poderosa na eficiência processual, na gestão documental e mesmo na equalização de acesso à justiça. Mas isso exige governança. É preciso que escritórios de advocacia, departamentos jurídicos e tribunais estabeleçam uma política clara para seu uso: quais sistemas podem ser utilizados, com que finalidade, por quem, em que condições de segurança, com qual grau de validação humana e com que tipo de registro e *accountability*. O CNJ deu um passo importante, nesse sentido, com a Resolução 615/2025.

Solução passa por três eixos

Em escritórios e ambientes privados, a realidade não é diferente. Muitos advogados utilizam ferramentas como ChatGPT, Copilot, Claude ou Gemini em atividades rotineiras, sem qualquer protocolo interno sobre confidencialidade, veracidade ou rastreabilidade. O uso pode ser inofensivo em uma revisão gramatical, mas torna-se problemático ao gerar um parecer técnico ou uma resposta automatizada a um cliente com base em conteúdo não auditável.

O que se configura, nesse contexto, é a sombra de uma prática digital paralela, cujos efeitos recaem sobre o número crescente de erros, distorções e insegurança jurídica. Além disso, a utilização de sistemas sem certificação ou sem infraestrutura adequada pode colocar os escritórios em situação de vulnerabilidade cibernética, sujeitando dados sensíveis de clientes a acessos indevidos, comprometendo a reputação institucional e a confiança na atuação profissional.

A solução passa por três eixos: educação, regulação e infraestrutura. Educação para que profissionais compreendam o funcionamento das ferramentas, seus limites e riscos. Regulação para estabelecer padrões mínimos de uso e responsabilidade. Infraestrutura para que os órgãos disponham de sistemas auditáveis, seguros e adaptados à realidade nacional. A governança da IA exige mecanismos internos claros: políticas formais de uso, mecanismos de consentimento, treinamentos obrigatórios, comitês internos de ética tecnológica e sistemas com rotulagem explícita de conteúdos gerados por IA.

Ferramentas utilizadas devem ser previamente avaliadas por equipes multidisciplinares quanto a segurança cibernética, impactos éticos, compatibilidade legal e viabilidade operacional. Escritórios e tribunais podem implementar ambientes controlados de experimentação (sandboxes), permitindo o uso monitorado da IA antes de sua aplicação plena. A governança deve incluir também planos de contingência em caso de mau funcionamento dos sistemas, diretrizes claras de responsabilização e rotinas de auditoria contínua.

Fragilidade da supervisão humana

No entanto, a realidade que se impõe é que, a cada dia que passa, Judiciário e advocacia promovem um “salto de fé” para o emprego da tecnologia, em uso generalizado e desprovido do mínimo de governança. Ademais, não podemos nos esquecer que se tornou comum ouvir a afirmação de que “tudo bem usar IA desde que haja supervisão humana”. Porém, essa ideia, embora intuitiva, é mais frágil do que parece.

Como explico há bastante tempo [6], há vários problemas de discriminação, erros e injustiças cometidos por sistemas de IA justamente sob supervisão humana, como no caso do COMPAS, software usado no sistema criminal norte-americano e que se mostrou prejudicial a pessoas pretas e latinas. Isso revela algo importante: colocar um humano para “verificar” decisões automatizadas não garante, por si só, segurança ou justiça. O humano também erra, tem vieses cognitivos, sofre pressões de tempo, toma atalhos mentais e pode tanto confiar demais na máquina quanto rejeitá-la quando ela contraria suas intuições ou preconceitos.

Ademais, uma pesquisa científica recente mostrou que o problema vai além dos limites psicológicos do ser humano. No estudo [7], pesquisadores analisaram por que as pessoas tendem a aceitar automaticamente a recomendação da IA mesmo quando sua própria resposta seria melhor. A descoberta central é simples, mas preocupante: isso acontece porque, na maioria dos sistemas, é racional confiar na máquina. Pensar dá trabalho; discordar exige esforço; revisar resultados demanda tempo. E, na prática, a pessoa recebe a mesma "recompensa" (ou seja, o mesmo benefício do acerto) tanto aceitando a sugestão da IA quanto pensando sozinha. Assim, seguir a máquina se torna a opção mais confortável e, infelizmente, a mais frequente.

O estudo demonstrou com clareza que ao comparar diferentes modos de recompensa, os pesquisadores viram que quando não há incentivo para pensar de forma independente, as pessoas praticamente deixam de questionar a IA. E quando se tentou dar um "bônus" para quem resolvesse sozinho, mas de forma igual para todas as situações, o resultado foi até pior: muitos participantes passaram a "fingir" que estavam resolvendo por conta própria apenas para ganhar o bônus, copiando exatamente a resposta da IA sem realizar esforço real; um exemplo clássico do que acontece quando um sistema mal desenhado incentiva comportamentos inadequados.

A solução mais eficiente encontrada pelos pesquisadores foi oferecer um bônus apenas quando a IA está incerta. Ou seja, incentivar o humano a pensar justamente quando ele tem mais chance de superar a máquina. Esse modelo, chamado de "bônus dinâmico", reduziu muito a dependência cega da IA e melhorou o desempenho geral das decisões. Na prática, quando vale mesmo a pena pensar, o sistema precisa sinalizar isso ao humano de maneira clara e recompensar esse esforço. Quando isso acontece, as pessoas passam a intervir com mais precisão.

Esse método produziu um efeito completamente diferente dos anteriores. Em vez de incentivar o oportunismo, como no bônus estático, o bônus dinâmico incentivou reflexão justamente onde ela era mais útil. As pessoas passaram a intervir de forma mais estratégica, pensavam quando a IA estava insegura e aceitavam a sugestão automatizada quando a IA demonstrava confiança.

O mais surpreendente é que esse modelo não apenas corrigiu o problema do excesso de confiança na IA, mas também gerou resultados mais eficientes e mais corretos. Ele fez com que o esforço humano fosse usado na hora certa, evitando desperdício de atenção e melhorando a qualidade final da decisão.

Letramento em IA, por si só, não resolve o problema

Esses achados são essenciais porque desmontam uma crença bastante difundida de que supervisão humana sempre protege contra erros. Supervisão só funciona se o ambiente induzir o humano a vigiar, revisar e questionar; e não a delegar. Se o sistema é construído de um modo que torna mais vantajoso aceitar automaticamente a IA, a supervisão se transforma em ficção. É como pedir que alguém revise um texto enquanto corre contra o relógio, sem tempo, sem incentivo para revisar e acreditando que o corretor automático raramente erra.

Ademais, outro estudo [8] mostra que, embora o letramento em IA seja importante, ele não resolve o principal problema para fins de supervisão humana: mesmo usuários com maior conhecimento técnico sobre IA tornam-se mais confiantes, porém não mais precisos ao avaliar o próprio desempenho. Em outras palavras, saber mais sobre IA não impede que a pessoa superestime sua capacidade nem melhora sua habilidade de identificar quando a resposta pode estar errada. Ou seja, mesmo quando a pessoa entende mais sobre como a IA funciona e sabe avaliar seus riscos, isso não significa que ela consiga supervisionar melhor o uso da ferramenta. Pelo contrário: quem tem mais "conhecimento técnico" e maior "capacidade crítica", tende a ficar mais confiante, mas não mais preciso ao julgar se acertou ou errou. Isso quer dizer que saber muito sobre IA não evita ilusões de segurança nem corrige erros de autoavaliação.

Por isso, assim como venho enfatizando, a supervisão humana precisa ser mais do que um slogan. Ela exige três pilares fundamentais: governança, condições adequadas de trabalho e literacia em IA.

Governança significa estabelecer regras institucionais claras: quando a IA decide, quando o humano deve revisar, quando a intervenção humana é obrigatória e como evitar que erros pequenos se tornem problemas gigantescos.

As condições adequadas de trabalho impõem a adoção de ferramentas customizadas para as tarefas empreendidas e seu monitoramento.

Finalmente, a literacia ou alfabetização em IA. Pessoas precisam entender que sistemas automatizados são probabilísticos, erram, podem discriminar e não possuem compreensão do mundo. Precisam aprender a reconhecer quando devem confiar

e quando devem duvidar. Sem essa compreensão, o humano pode apenas apertar “OK” sem perceber os riscos ou rejeitar a IA por puro preconceito tecnológico.

Conclusão

Em síntese, a supervisão humana só funciona quando o sistema é projetado para ajudar o humano a pensar, e não para substituí-lo silenciosamente. Incentivos adequados, transparência, calibragem da confiança, monitoramento constante e educação digital são os elementos que fazem a supervisão deixar de ser uma ilusão regulatória e se tornar uma proteção real.

Perceba-se, por derradeiro, que não se trata de conter a inovação, mas de discipliná-la. *Shadow AI* é o sintoma de uma adoção tecnológica sem maturidade institucional. O desafio é dissipar as sombras com conhecimento, normatividade e responsabilidade. É urgente que instituições jurídicas invistam em programas de capacitação contínua sobre IA, em comitês que permitam rastrear, auditar e corrigir os usos indevidos. A governança de IA [9] não é um luxo ou tendência, mas um imperativo para proteger direitos fundamentais, garantir a integridade do sistema de justiça e preservar os valores democráticos diante da disrupção tecnológica em curso.

---

[1] NUNES. IA generativa no Judiciário: como evitar muito gasto e pouco resultado. Aqui NUNES. Implantação da IA no Judiciário passa ao largo do debate público. Aqui

[2] NUNES. IA no Judiciário. Aqui

[3] De acordo com o levantamento da Komprise no âmbito empresarial, com mais de 200 diretores de TI, aproximadamente 90% dos gestores demonstram preocupação com a Shadow AI, e 79% afirmaram já ter enfrentado efeitos adversos decorrentes do uso não autorizado de IA por colaboradores, incluindo exposição indevida de dados pessoais e geração de informações incorretas: aqui

[4] NUNES; DUTRA, V. Workslop no Direito: o custo gerado pelo uso desleixado da IA no Direito. Aqui

[5] CNJ. *Pesquisa Inteligência Artificial no Judiciário 2024*: aqui

[6] NUNES, D. A supervisão humana das decisões de inteligência artificial reduz os riscos? Aqui

[7] HOLSTEIN, J. et al. When Thinking Pays Off: Incentive Alignment for Human-AI Collaboration. Aqui

[8] FERNANDES, D; et al. *AI makes you smarter but none the wiser: The disconnect between performance and metacognition*. Aqui

[9] NUNES. Implantação da IA no Judiciário passa ao largo do debate público. Aqui

Fonte: https://conjur.jumps.com.br/2025-dez-16/shadow-ai-no-direito-o-risco-do-uso-da-ia-nas-sombras/