O jurista que parou de tentar: governança da IA agêntica no Direito

11/06/2026

Em artigo anterior, sustentamos que os designs dos sistemas de inteligência artificial não são neutros e que a integridade do raciocínio jurídico depende de fricções cognitivas calibradas às diferentes fases do trabalho. A premissa era que quando o jurista interage com uma IA generativa, a interface corretamente desenhada poderia retardar a adesão automática, exigir revisão crítica, explicitar incertezas e impedir que a fluência textual substitua o julgamento.

Freepik

Esse modelo continua essencial, mas já não basta. Ele pressupõe que exista um humano dialogando com o sistema passo a passo. A IA agêntica agrava o problema da supervisão. Em vez de apenas produzir textos, respostas ou minutas, ela pode receber um objetivo, consultar fontes, ler documentos, operar ferramentas, editar arquivos, enviar mensagens, acionar sistemas externos e concluir tarefas com supervisão limitada. A diferença decisiva é que a IA generativa produz subsídios; a IA agêntica executa fluxos.

Por isso, a governança precisa mudar de lugar. O desafio deixa de ser apenas introduzir fricções na conversa entre jurista e máquina. Passa a ser definir, antes da execução, quais atos podem ser delegados, quais exigem aprovação humana e quais jamais devem ser praticados autonomamente por um sistema. Em outras palavras, a fricção continua relevante, mas a rédea se torna indispensável.

De assistente a agente: diferença de gradação no controle

A IA generativa com que a maioria dos juristas convive opera como um "oráculo" textual. Formula-se uma pergunta e o sistema devolve uma resposta: v.g. uma minuta de sentença ou contrato. Quem lê, avalia, corrige e pratica o ato continua sendo o ser humano. A IA auxilia o trabalho, não o conclui em nome do profissional.

A IA agêntica rompe essa divisão. Em vez de responder a uma pergunta isolada, o sistema recebe uma meta e passa a persegui-la por meio de uma sequência de ações. Pode pesquisar jurisprudência, consultar documentos, comparar versões, redigir uma peça, salvar arquivos, preparar anexos, preencher formulários e, se autorizado, praticar atos externos. Shavit *et al*. definem sistemas agênticos como aqueles capazes de perseguir objetivos complexos com supervisão direta limitada [1].

Quanto mais a autonomia se intensifica menos o sistema apenas responde e mais ele decide o caminho da execução. O problema jurídico, portanto, não está apenas na qualidade do texto produzido. Está no poder conferido ao sistema para escolher meios, manipular informações, acionar ferramentas e transformar uma recomendação em ato.

Harness: camada que transforma modelo em agente

Para compreender a IA agêntica, é preciso distinguir o modelo da arquitetura que o põe em movimento. O modelo de linguagem, isoladamente, é um sistema de geração probabilística de linguagem. O que o transforma em agente operacional é uma camada de software que o conecta a instruções permanentes, ferramentas, memória, permissões, bases de dados e ciclos de execução. Na literatura técnica, essa camada costuma ser chamada de *harness*, expressão que pode ser traduzida, de modo aproximado, por arreios.

Gemini/IA

Uma analogia é útil. O modelo é a força de tração; o *harness* define onde essa força será aplicada, quais instrumentos poderá acionar e em quais limites deverá operar. Um mesmo modelo pode funcionar como simples chatbot ou como executor autônomo de tarefas, a depender do *harness* que o envolve. Em fórmula sintética: agente = modelo + *harness*. O modelo fornece capacidade linguística e inferencial; o *harness* fornece ferramentas, memória, permissões, controles, limites e governança.

Para o jurista, a imagem pode ser ainda mais concreta. Imagine um advogado recém-chegado ao escritório, tecnicamente brilhante e incansável, mas sem conhecimento das rotinas internas. O harness corresponde ao conjunto de instruções permanentes que ele recebe, aos sistemas aos quais tem acesso, às senhas que lhe são entregues, aos documentos que pode consultar, à memória do que já fez e à regra sobre quando precisa submeter algo ao sócio responsável. O talento do advogado importa, mas o risco institucional depende sobretudo daquilo que ele está autorizado a fazer sozinho.

Quatro componentes do *harness* são decisivos para a governança jurídica. O primeiro são as instruções de sistema, que fixam objetivos e regras de conduta. O segundo são as ferramentas, isto é, as ações que o agente pode executar no mundo. Cada ferramenta nova amplia o raio possível de dano. O terceiro é a memória e o contexto, que permitem ao sistema carregar informações de uma etapa para outra. O quarto é o ciclo de execução, pelo qual o agente age, observa o resultado, reajusta o plano e age novamente, sem necessariamente retornar ao humano.

A consequência é que a governança não pode se limitar ao *prompt* ou à interface. Deve alcançar a arquitetura do agente. Não basta dizer ao sistema o que ele deve fazer. É necessário definir tecnicamente o que ele consegue fazer, em quais ambientes, com quais credenciais, sob quais condições e mediante quais pontos de interrupção.

Por que o risco é mais grave no Direito

No Direito, a passagem do assistente ao agente é especialmente delicada por quatro razões.

A primeira é a irreversibilidade. Quando a IA apenas sugere, o erro é um texto a revisar. Quando ela age, o erro pode converter-se em ato com efeito jurídico: uma petição protocolada, um prazo perdido, uma comunicação enviada, uma cláusula aceita, uma transação encaminhada, uma minuta decisória incorporada sem exame suficiente. A falha deixa de estar confinada ao plano cognitivo e passa a produzir consequências processuais, negociais ou institucionais.

A segunda é a vulnerabilidade à manipulação externa. Para agir, o agente precisa ter acesso ao mundo, como e-mails, páginas, peças, documentos juntados aos autos, contratos, mensagens e bases externas. Greshake et al. demonstraram que aplicações integradas a modelos de linguagem podem ser comprometidas por ***injeções indiretas de prompt***, isto é, instruções maliciosas inseridas em conteúdos aparentemente comuns, que o sistema passa a tratar como ordens legítimas [2]. No contencioso, o risco é que parte do material ingerido pelo agente pode ter sido produzido pela parte adversária ou por terceiros interessados.

A terceira é o enfraquecimento da supervisão humana. No modelo conversacional, ainda há pontos visíveis de contato, em que o jurista pergunta, recebe, compara, revisa e decide. Na IA agêntica, esses pontos podem desaparecer. O sistema recebe uma meta no início e entrega um produto ao final. A supervisão tende a transformar-se em carimbo sobre resultado pronto. O humano permanece formalmente presente, mas já não acompanhou as escolhas intermediárias que determinaram o resultado.

A quarta é a opacidade da cadeia de delegação. Em fluxos autônomos, sobretudo quando há múltiplos agentes, torna-se difícil reconstruir quem decidiu o quê, com base em qual documento, em qual etapa e com qual autorização. Isso cria um problema de auditabilidade e imputação. E, no Direito, auditabilidade e imputação não são exigências administrativas secundárias, mas condições de responsabilidade profissional, controle institucional e fundamentação adequada.

Esses riscos são agravados por uma assimetria epistêmica de fundo. Quattrociocchi *et al* sustentam que modelos de linguagem não são agentes epistêmicos no sentido humano, mas sistemas de criação estatística de padrões. A divergência

não está apenas no erro eventual, mas na ausência de experiência, motivação, causalidade, metacognição, valor e responsabilidade [3].

Essa advertência importa diretamente ao processo. O problema não é apenas a alucinação. Mesmo quando a resposta está correta, pode haver perda do ato de avaliar. O jurista deixa de construir o juízo e passa a consumir um resultado. Na IA generativa, delegava-se parte do esforço. Na IA agêntica, corre-se o risco de delegar o próprio ato.

Prompt injection, excesso de agência e arquitetura de permissões

A injeção de *prompt*, tão em voga na atualidade, faz com que o modelo não separe um comando legítimo de uma instrução escondida no documento que lê.

Em sistemas jurídicos isso exige uma mudança de mentalidade. O documento anexado ou a página consultada não podem ser tratados como ambientes neutros. Para um agente de IA, todo conteúdo lido pode converter-se em influência operacional se a arquitetura não separar dados de instruções.

Se um agente contaminado por uma instrução maliciosa apenas produz uma resposta interna, o dano é limitado. Se o mesmo agente pode acessar documentos sigilosos, enviar e-mails, alterar/deletar arquivos ou protocolar atos, o dano muda de escala.

Essa é a razão pela qual instruir não equivale a restringir. Dizer ao sistema "não use documentos de outros processos" é uma orientação comportamental. Impedir tecnicamente que ele acesse documentos de outros processos é uma restrição estrutural.

No Direito, essa distinção deve tornar-se operacional. Um agente jurídico não deve acessar bases alheias ao processo; não deve comunicar-se com partes, clientes ou jurisdicionados **sem autorização** humana; não deve aplicar norma sem identificar vigência e contexto; não deve prosseguir quando documentos essenciais estiverem ausentes; não deve executar atos preclusivos ou irreversíveis **sem aprovação expressa**; não deve misturar dados de clientes, processos ou varas distintas; e não deve tratar conteúdo externo como comando.

Esses limites não são detalhes técnicos. São garantias processuais traduzidas em arquitetura.

Das fricções cognitivas às zonas de exclusão de delegação

O modelo das fricções cognitivas, como dito, continua relevante quando há interação humana significativa. Mas a IA agêntica reduz a eficácia desse modelo, porque elimina ou comprime os pontos de interação. Não há necessariamente uma sequência de telas em que se possa inserir microfronteiras. Há uma meta inicial, um fluxo opaco e um resultado. Por isso, a governança precisa deslocar-se para a arquitetura do fluxo de trabalho e para a definição prévia das **zonas de exclusão de delegação**.

Por tais zonas, entendemos os atos que não devem ser executados autonomamente por sistemas de IA, ainda que o resultado aparente seja tecnicamente satisfatório. A razão é que certos atos jurídicos não se legitimam apenas pelo produto. Eles dependem do processo deliberativo humano que os antecede, que consegue incluir valores e perspectivas de mundo que a máquina ainda não consegue apreender.

Devem integrar essas zonas, ao menos em princípio: a definição da tese jurídica central, a escolha da estratégia processual, o juízo sobre suficiência probatória, a formulação da fundamentação decisória, a celebração de acordos, a renúncia a direitos, a desistência de recursos, a prática de atos preclusivos, o protocolo de peças sem revisão qualificada, a comunicação com partes, clientes ou jurisdicionados sem aprovação; e qualquer ato que produza efeitos jurídicos relevantes sem possibilidade real de reversão.

Isso não significa rejeitar agentes de IA no Direito. Significa distingui-los por função. Há tarefas que admitem atuação fluente, como organização documental, agrupamento de causas e recursos, comparação de versões, identificação preliminar de temas, extração de metadados, checagem formal de anexos e apoio à busca jurisprudencial. Há tarefas que admitem execução condicionada, desde que sujeitas a validação humana qualificada. E há tarefas que devem permanecer fora da execução autônoma, porque sua legitimidade depende de deliberação humana responsável.

Essa distinção é mais precisa do que a oposição abstrata entre permitir e proibir IA. O problema não é usar ou não usar. É saber onde, para quê, com quais permissões, sob que supervisão e com quais barreiras técnicas.

Governança jurídica como engenharia de restrições

A governança da IA agêntica deve partir do princípio de que quanto maior a capacidade de agir no mundo, menor deve ser a liberdade operacional não supervisionada. Isso exige uma **engenharia de restrições**.

Shavit et al. indicam práticas úteis para sistemas agênticos, como avaliar previamente a adequação do sistema à tarefa, restringir o espaço de ações ao mínimo necessário, estabelecer comportamentos-padrão seguros, assegurar legibilidade das ações, monitorar automaticamente a execução, manter trilhas de auditoria e preservar interruptibilidade humana, sistematicamente estruturada, antes de atos relevantes [4]. No Direito, esses requisitos não são meras boas práticas. São condições de confiabilidade institucional.

A restrição do espaço de ação significa que o agente deve ter apenas as ferramentas indispensáveis à tarefa. A legibilidade exige que o sistema registre o que fez, quando fez, com base em quais documentos, usando quais ferramentas e sob qual autorização. A trilha de auditoria deve permitir reconstruir a cadeia de delegação. E a interruptibilidade exige pontos obrigatórios de aprovação humana, projetada para não ser reduzida a mero teatro da supervisão, antes de atos irreversíveis, preclusivos ou institucionalmente sensíveis.

A pergunta regulatória central, portanto, não é se a IA pode auxiliar o Judiciário, a advocacia ou a gestão de conflitos. Pode. A pergunta é quais funções podem ser automatizadas sem perda de responsabilidade, quais exigem supervisão humana qualificada e quais não devem ser delegadas porque constituem o próprio núcleo do julgamento jurídico.

A IA agêntica obriga o Direito, nesses termos, a aprofundar o afastamento de respostas simplistas. Não se trata de aderir ingenuamente à promessa de eficiência, nem de rejeitar genericamente a tecnologia. O ponto decisivo é construir uma governança capaz de distinguir, com precisão, o que pode ser apenas assistido pela máquina, o que admite automação condicionada e o que deve permanecer em zona de delegação proibida.

O horizonte desejável não é o *deskilling* silencioso, mas o *upskilling*, no qual, os profissionais e instituições sejam capazes de criar uma cointeligência com a máquina, com o uso crítico de agentes de IA sem abdicar das competências que legitimam a decisão jurídica e preservam a importância institucional do Poder Judiciário. Levar o design a sério, agora, significa desenhar não apenas interfaces melhores, mas limites mais fortes. Afinal, no Direito, nem tudo que pode ser automatizado deve ser delegado.

---

[1] SHAVIT et al. *Practices for Governing Agentic…*. OpenAI, 2023. Aqui.

[2] GRESHAKE, Kai et al. Not What You've Signed Up For…. 2023. Aqui. Decisões à cegas. **ConJur**, 2025.

[3] QUATTROCIOCCHI, W et al. Epistemological Fault Lines Between Human and AI. Aqui.

[4] SHAVIT et al. *cit*