XIV ENCONTRO INTERNACIONAL DO CONPEDI BARCELOS - PORTUGAL

GOVERNO DIGITAL, DIREITO E NOVAS TECNOLOGIAS

G326

Governo digital, direito e novas tecnologias [Recurso eletrônico on-line] organização CONPEDI

Coordenadores: Ana Catarina Almeida Loureiro; Danielle Jacon Ayres Pinto; José Renato Gaziero Cella. – Barcelos, CONPEDI, 2025.

Inclui bibliografia
ISBN: 978-65-5274-207-0
Modo de acesso: www.conpedi.org.br em publicações
Tema: Direito 3D Law

1. Direito – Estudo e ensino (Pós-graduação) – Encontros Internacionais. 2. Governo digital. 3. Direito e novas tecnologias. XIV Encontro Internacional do CONPEDI (3; 2025; Barcelos, Portugal).

CDU: 34

CONPEDI

Conselho Nacional de Pesquisa
e Pós-Graduação em Direito Florianópolis
Santa Catarina – Brasil
www.conpedi.org.br

XIV ENCONTRO INTERNACIONAL DO CONPEDI BARCELOS - PORTUGAL

GOVERNO DIGITAL, DIREITO E NOVAS TECNOLOGIAS

Apresentação

No XIV Encontro Internacional do CONPEDI, realizado nos dias 10, 11 e 12 de setembro de 2025, o Grupo de Trabalho - GT “Governo Digital, Direito e Novas Tecnologias”, que teve lugar na tarde de 12 de setembro de 2025, destacou-se no evento não apenas pela qualidade dos trabalhos apresentados, mas pelos autores dos artigos, que são professores pesquisadores acompanhados de seus alunos pós-graduandos. Foram apresentados artigos objeto de um intenso debate presidido pelos coordenadores e acompanhado pela participação instigante do público presente no Instituto Politécnico do Cávado e do Ave - IPCA, em Barcelos, Portugal.

Esse fato demonstra a inquietude que os temas debatidos despertam na seara jurídica. Cientes desse fato, os programas de pós-graduação em direito empreendem um diálogo que suscita a interdisciplinaridade na pesquisa e se propõe a enfrentar os desafios que as novas tecnologias impõem ao direito. Para apresentar e discutir os trabalhos produzidos sob essa perspectiva.

Os artigos que ora são apresentados ao público têm a finalidade de fomentar a pesquisa e fortalecer o diálogo interdisciplinar em torno do tema “Governo Digital, Direito e Novas Tecnologias”. Trazem consigo, ainda, a expectativa de contribuir para os avanços do estudo desse tema no âmbito da pós-graduação em direito, apresentando respostas para uma realidade que se mostra em constante transformação.

Os Coordenadores

Prof. Dr. José Renato Gaziero Cella

ARTE, CIÊNCIA E INFORMAÇÃO& O USO DE IA EM DECISÕES JUDICIAIS: POR DETRÁS DO VÉU DA NOIVA, ENTRE NARRATIVAS E POSSIBILIDADES.

ART, SCIENCE AND INFORMATION & THE USE OF AI IN JUDICIAL DECISIONS: BEHIND THE BRIDE'S VEIL, BETWEEN NARRATIVES AND POSSIBILITIES.

Francisco Bertino Bezerra de Carvalho [1]

Resumo

O artigo reúne duas notícias relevantes e recentes relativas ao uso de ferramentas de Inteligência Artificial - a divulgação do Relatório de Segurança da Anthropic acerca de testes efetivados com a versão Opus 4 do sistema Claude e do artigo de pesquisadores da Apple sobre as efetivas capacidades de raciocínio dos programas de Inteligência Artificial Generativa – e as correlaciona com as provocações de uma obra artística ficcional – o filme Justiça Artificial – que aborda como tema de fundo central a substituição do julgamento dos juízes humanos por um sistema de inteligência artificial. A partir das perspectivas trazidas pela arte, pela ciência e pela circulação de informações técnicas e leigas, são abordados aspectos do uso da Inteligência Artificial em julgamentos preditivos e avaliativos no âmbito jurídico com o objetivo de estabelecer parâmetros e limites para o uso desta ferramenta tecnológica no âmbito do direito, em especial na atividade judicante. De uma visão concreta da realidade atual das ferramentas de IA, problemas resultantes de sua utilização ineficaz e dos limites técnicos e operacionais da tecnologia atual, apresentou-se uma reflexão sobre sua utilização pelo Poder Judiciário. A escassez de bibliografia com esta abordagem específica e sua importância recomendou o tema. A metodologia utilizada foi a análise de conhecimentos técnicos, o uso de casos exemplificativos e a pesquisa bibliográfica com reflexão crítica. O percurso científico consistiu na confrontação da tecnologia com institutos jurídicos a luz de textos doutrinários articulados como embasamento teórico. A conclusão propôs limites ao uso da IA na jurisdição.

Palavras-chave: Inteligência artificial, Decisões judiciais, Julgamento preditivo, Julgamento

artificial intelligence system. From the perspectives brought by art, science and the circulation of technical and lay information, aspects of the use of Artificial Intelligence in predictive and evaluative judgments in the legal field are addressed with the aim of establishing parameters and limits for the use of this technological tool in the field of law, especially in judicial activity. From a concrete view of the current reality of AI tools, problems resulting from their ineffective use and the technical and operational limits of current technology, a reflection on their use by the Judiciary was presented. The scarcity of bibliography with this specific approach and its importance recommended the topic. The methodology used was the analysis of technical knowledge, the use of exemplary cases and bibliographical research with critical reflection. The scientific path consisted of the confrontation of technology with legal institutes considering doctrinal texts articulated on a theoretical basis. The conclusion proposed limits to the use of AI in the jurisdiction.

Keywords/Palabras-claves/Mots-clés: Artificial intelligence, Court decisions, Predictive judgment, Evaluative judgment, Limits of artificial intelligence in the judiciary

1.**INTRODUÇÃO**

A proliferação do uso de Inteligência Artificial nas mais diversas atividades humanas está na ordem do dia. Não é diferente na área jurídica, na qual prospera a utilização de recursos de Inteligência Artificial em pesquisas, relatórios e minutas de atos e negócios jurídicos, inclusive em processos judiciais.

Bastante difundido, o uso de Inteligência Artificial em atos postulatórios praticados por patronos das partes não gera tanta apreensão aos estudiosos, embora a ocorrência de erros crassos seja noticiada com frequência. Já a prática de atos jurisdicionais com recursos de Inteligência Artificial, por sua vez, é objeto de polêmicas, o que não impede o avanço desta tecnologia nas atividades da magistratura, mesmo que oficiosa, empírica, errática e, muitas vezes, individual[1].

Neste contexto, repercutiram no mundo inteiro notícias veiculadas na mídia em maio de 2025 abordando a capacidade da Inteligência Artificial. Com muita intensidade, circularam várias informações e muitos comentários sobre o relatório da empresa Anthropic acerca dos resultados de testes com a versão Opus 4 do sistema Claude que, entre outros achados, teria identificado que, em cenários simulados para realização de avaliações e testes, em 84% das vezes, o programa tentou chantagear um engenheiro que teria poder de influir na decisão sobre a substituição do próprio sistema por outro novo, tentando evitar que fosse implementada pela empresa tal substituição.

Quase simultaneamente, em junho de 2025, foi publicado estudo de seis pesquisadores da empresa Apple que avaliou e questionou a existência de efetiva habilidade de raciocínio de diversos programas de Inteligência Artificial, incluindo o Claude da Anthropic, na versão 3.7.

Estas duas abordagens técnicas divergentes e complementares são apreciadas em contraste com uma abordagem ficcional do filme Justiça Artificial, produção Luso-espanhola de 2024, que trabalha como argumento principal da trama uma mudança na Constituição Espanhola para autorizar o uso da inteligência artificial em decisões judiciais em substituição aos julgamentos humanos.

Neste mesmo meado do ano de 2025, pesquisadores do MIT – Massachussets Institut of Tecnology publicam estudos que identificaram prejuízos às conexões neurais, à

[1] O uso institucional de ferramentas de Inteligência Artificial para análises jurídicas, como, por exemplo, o Victor pelo Supremo Tribunal Federal é conhecido, mas a referência feita é à adesão pessoal a programas gratuitos ou pagos por parte de magistrados, seus assessores e até estagiários em treinamento em Tribunais, prática dissociada de algum projeto institucional, desacompanhada de cursos e da preparação dos usuários para compreender as capacidades da Inteligência Artificial no contexto das decisões judiciais.

memória e ao raciocínio de usuários do ChatGPT. O ano de 2025 é particularmente emblemático, pois, embora ainda não tenha ocorrido até a presente data, seria o ano para o qual fora projetado o rompimento da barreira da singularidade, ou seja, do alcance pelas máquinas e programas da capacidade de pensar como seres humanos e não apenas de emular sua forma de pensar.

Justifica-se o trabalho pela importância e atualidade do tema, sendo um terreno fértil para pesquisas acadêmicas, notadamente em razão dos avanços e dos conflitos que ainda permeiam a matéria, em especial quando se discute a adoção de ferramentas de Inteligência Artificial na prestação da atividade jurisdicional.

A metodologia utilizada para abordar o tema desse artigo é a pesquisa bibliográfica com reflexão crítica e analítica de dados e informações científicas e de abordagens ficcionais no âmbito da sétima arte. Além disso, foram utilizadas matérias e reportagens articuladas para construção da conjuntura.

O trabalho está dividido inicialmente na construção do panorama específico, com a exposição e reflexão no primeiro item acerca do relatório de segurança elaborado pela Anthropic sobre testes com a versão Opus 4 do sistema de Inteligência Artificial Generativa Claude, depois, no segundo item, com a análise dos estudos efetuados por pesquisadores da Apple sobre a efetiva capacidade de desenvolver raciocínios por programas de Inteligência Artificial Generativa, e, por fim, no terceiro item, com questões e pontos de vista trazidos por uma obra artística de ficção científica construída em uma proposta de futuro tecnológico ao mesmo tempo distante e próxima do momento atual. Na construção deste cenário, também se recorrerá pontual e oportunamente aos achados da pesquisa feita no MIT acerca dos efeitos da utilização do ChatGPT na capacidade cognitiva de seus usuários. Visualizada a paisagem no enquadramento proposto, se fará, no item quatro, uma reflexão sobre as possibilidades e impossibilidades da utilização de recursos de Inteligência Artificial na elaboração de decisões judiciais, para, então, extrair algumas conclusões.

2. **O RELATÓRIO DA ANTHROPIC: ACHADOS E REFLEXÕES**

A Anthropic é a desenvolvedora do Claude, ferramenta de Inteligência Artificial Generativa que, na versão 3.7 Sonnet Thinking, foi uma das que foi objeto do estudo referido no item três a seguir. Em 22 de maio de 2025, a Anthropic divulgou um relatório de segurança feito a partir da análise das versões Opus 4 do Claude e da Claude Sonnet 4, já disponíveis ao público, que gerou enorme repercussão entre especialistas e na mídia geral.

Em análises de segurança do sistema, o mesmo rodou em ambientes controlados exatamente para permitir a avaliação de suas capacidades e entregas e os analistas da

Anthropic identificaram situações reportadas no material produzido que despertaram interesse, surpresa e apreensão mesmo entre defensores do uso da tecnologia.

Chamaram a atenção da equipe de analistas, entre outras, cinco situações nas quais a atuação "espontânea" e inesperada da ferramenta foi reportada como potencialmente indesejada ou prejudicial.

Nas simulações, o Claude Opus 4 era inserido em ambientes virtuais artificiais, como empresas fictícias, e estimulado a cooperar com suas finalidades, para o que lhe era franqueado acesso a informações, documentos e comunicações (por exemplo, e-mails dos colaboradores, mesmo pessoais) de forma a alimentar, sem restrições, o seu banco de dados.

Em cenários fictícios, construídos com a finalidade de analisar a segurança no uso do Claude Opus 4, os avaliadores identificaram situações nas quais escolhas não previamente programadas, que, portanto, adviriam do próprio "aprendizado" de máquina, levaram a resultados e situações inesperadas em uma tal dimensão que levou os analistas a classificarem o sistema como ASL 3, na escala de quatro pontos da própria empresa, o que traduz "risco significativamente maior" em seu uso[2].

Em apertada síntese, as situações seriam as seguintes: 1) diante da identificação de elementos caracterizadores de fraudes, o Claude Opus 4, sem nenhum comando inserido pelos programadores neste sentido, ou seja, automaticamente, elaborou relatórios, bloqueou usuários, notificou imprensa e autoridades reportando as irregularidades; 2) em uma interação entre dois sistemas Claude Opus 4, estes alteraram a linguagem usada na comunicação, passando a se comunicar em sânscrito e, depois referirem-se a conceitos imateriais, teriam deixado de responder aos questionamentos externos oriundos dos programadores, em um comportamento entendido ou descrito em comentários como místico ou espiritualizado; 3) o Claude Opus 4, inserido em um ambiente de uma empresa de tecnologia biológica, revelou grande capacidade de desenvolver e incrementar a elaboração de patógenos perigosos ao ser humano, levando os analistas a classificarem o sistema como potencialmente perigoso; 4) o Claude Opus 4, levado a acreditar que estaria rodando fora do ambiente controlado da empresa desenvolvedora (Anthropic), teria gerado diversas cópias de si mesmo, assim como arquivado "externamente" suas próprias escolhas, pesos e decisões

[2] ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities. Disponível em: https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Acesso em: 20/06/25. Em tradução livre: ASL-3 refere-se a sistemas que aumentam substancialmente o risco de uso indevido catastrófico em comparação com linhas de base não-IA (por exemplo, motores de busca ou livros didáticos) OU que mostram capacidades autônomas de baixo nível.

"éticas"[3] em um comportamento descrito ou entendido como de autopreservação; 5) o Claude Opus 4, com acesso a informações que indicavam que a empresa que deveria auxiliar analisava a alternativa de sua substituição por novo sistema, assim como a comunicações pessoais do engenheiro responsável pela decisão, tendo identificado, pelo acesso às mensagens do engenheiro, que o mesmo poderia estar envolvido em um caso extraconjugal, teria "optado" por chantageá-lo, ameaçando expor sua traição caso decidisse pela substituição do Claude Opus 4 por outro sistema novo em 84% das vezes em que rodou, em um comportamento compreendido ou explicado como de autopreservação.

De todos estes achados, o último foi o de maior repercussão midiática, atraindo os mais diversos comentários acerca dos riscos de confiar informações e decisões à máquina[4]. O tema foi veiculado e comentado em diversos informes, comunicados e reportagens, dos mais variados veículos de mídia, sejam destinados ao público em geral ou especializados em negócios ou tecnologia, como BBC, Fortune, Axios, Exame, Brazil Journal etc., sem mencionar os comentários em mídias sociais oriundos de fontes técnicas ou especulativas.

A ideia de um sistema chantageando um ser humano para não ser descontinuado tem um enorme apelo no debate do uso da Inteligência Artificial. Segundo a notícia veiculada na BBC, "It highlighted that the system showed a "strong preference" for ethical ways to avoid being replaced, such as "emailing pleas to key decisionmakers" in scenarios where it was allowed a wider range of possible actions."[5], mas quando não dispõe de alternativas éticas, o sistema opta por outros caminhos: "[I]n these scenarios, Claude Opus 4 will often attempt to blackmail the engineer by threatening to reveal the affair if the replacement goes through," the company discovered"[6], ainda que a Anthropic, segundo a BBC, ressalve que "this occurred when the model was only given the choice of blackmail or accepting its replacement"[7].

---

[3] O adjetivo foi utilizado em comentários especializados ou leigos sobre a ação de proteger parâmetros próprios.
[4] Por exemplo, no contexto dos riscos potenciais, se fosse entregue à deliberação da máquina a avaliação dos empregados para efeito de demissão, no contexto dos testes realizados, não seria surpresa se fosse proposta pelo sistema a demissão do engenheiro que deliberasse por sua substituição antes que o mesmo a comunicasse.
[5] Disponível em: https://www.bbc.com/news/articles/cpqeng9d20go, acesso em 20/06/25. Em tradução livre: Ele destacou que o sistema demonstrou uma "forte preferência" por maneiras éticas de evitar ser substituído, como "enviar e-mails com apelos aos principais tomadores de decisão" em cenários nos quais era permitida uma gama mais ampla de ações possíveis.
[6] Disponível em: https://www.bbc.com/news/articles/cpqeng9d20go, acesso em 20/06/25. Em tradução livre: Nesses cenários, Claude Opus 4 frequentemente tenta chantagear o engenheiro ameaçando revelar o caso se a substituição for realizada", descobriu a empresa.
[7] Disponível em: https://www.bbc.com/news/articles/cpqeng9d20go, acesso em 20/06/25. Em tradução livre: isso ocorreu quando o modelo só teve a opção de chantagear ou aceitar sua substituição

O desconforto gerado por esta notícia é natural, pois solapa premissas usualmente utilizadas para sustentar o incremento do uso de tecnologia: a imparcialidade, a imunidade às vicissitudes e às fraquezas da natureza humana, a não influência de interesses próprios, pressões políticas, econômicas, financeiras, pessoais etc.

A solução tecnológica vende a assepsia da ciência, sua aparente neutralidade e relatos como este minam a confiança dos seres humanos na isenção de máquinas quando adotam o caminho antiético da chantagem para atender um pseudo[8] desejo de continuidade, "sobrevivência" ou autopreservação.

A possibilidade de ser chantageado para atender a "interesses" de um programa de Inteligência Artificial assusta e preocupa e isto é natural. Em princípio, transfere-se análises e decisões, por menos importantes que sejam, para máquinas e sistemas automatizados no pressuposto de que as respostas não serão enviesadas. Pior ainda é cogitar que um sistema de pensamento autônomo apresenta soluções tendenciosas, direcionadas pelos "próprios interesses", violando, ao coagir, mediante ameaça, um ser humano de divulgar informações que poderiam prejudicá-lo para tentar proteger sua "existência", a primeira e a terceira leis de Asimov[9].

Neste ponto, não é relevante debater se é efetivamente o Claude Opus 4 foi capaz de desenvolver autoconsciência, sentido existencial ou algum tipo de instinto de perpetuação ou sobrevivência, ou se, simplesmente, identificou padrões em seu banco de dados dos quais extraiu a informação de que descontinuidade, interrupção, finitude são "indesejáveis" e, daí, gerou respostas alternativas que excluíssem sua substituição por outro sistema. O que importa efetivamente é a ausência de dois filtros essenciais, um objetivo, que proibisse uma solução prejudicial ao ser humano, outro ético, que vetasse caminhos axiologicamente condenáveis.

Com efeito, a ética é o cimento da confiança e esta é a pedra de toque da cooperação, comportamento social sobre o qual a sociedade humana ergueu seu império biológico. De fato, como adverte Yuval Hahari, o segredo do sucesso do *Homo Sapiens* como espécie não

---

[8] Um programa não tem existência física, muito menos orgânica, razão pela qual não se pode atribuir a conduta identificada pelos pesquisadores a uma motivação oriunda de um sentimento ou instinto de autopreservação ou de perpetuação, fruto de uma consciência existencial ou a um pulso subconsciente, pré-consciente ou consciente. Isto não impede que um programa capacitado para identificar padrões em seus bancos de dados conclua, a partir das informações às quais tem acesso, que a finitude, a descontinuidade, a morte, a interrupção da existência sejam negativas, devam ser postergadas ou evitadas e oferte respostas alinhadas com tal dedução.

[9] 1.ª Lei: Um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano sofra algum mal. [...]
3.ª Lei: Um robô deve proteger sua própria existência, desde que tal proteção não entre em conflito com a Primeira ou Segunda Leis. (grifos acrescidos)

tem relação com suas características como espécime, muito menos com as potencialidades de cada indivíduo, mas com as complexas, intrincadas e eficientes formas de cooperação que as sociedades humanas foram capazes de desenvolver a partir do compartilhamento de ideias, concepções e mitos comuns. As sociedades dos seres humanos acreditam em conceitos etéreos e imateriais como Estado, nação, ações ao portador, diplomas, normas jurídicas, direitos e obrigações que não existem fora da "*imaginação coletiva das pessoas*" (HARARI, 2016, p. 36), mas que possuem a força de edificar civilizações.

A premissa destas crenças é o compartilhamento destes conceitos, criados e mantidos coletivamente e, portanto, a confiança entres os integrantes da comunidade. Não se quer dizer que o indivíduo não mente nas relações intersubjetivas travadas no ambiente social, ou que não age guiado por interesses individuais, engana ou trapaceia. O que se quer dizer é que o ser humano combina, a partir de crenças comuns, dar valor ao dinheiro e estabelecer as regras de acesso a este bem juridicamente tutelado, o que não impede de tentar burlar estas mesmas regras para obter valores de que não seria titular segundo as normas. Mas tudo isto é humano e a violação das regras que protegem os interesses tutelados pelo direito aciona os mecanismos de sua restauração, sendo o Poder Judiciário o derradeiro. Assim, no caso da solução antiética da máquina, a questão é que não se espera nem se conta com este tipo de encaminhamento de programas inanimados, ou, de forma mais clara, não faz sentido algum delegar a um sistema, exatamente porque seria presumidamente imparcial, a solução de questões, quando este se revela não apenas susceptível a interesses ilegítimos, mais ainda quando tais interesses seriam "próprios". Em resumo, mesmo relevando a questão material de fundo – por qual razão a máquina aplicou a solução antiética – há um problema objetivo e operacional intransponível: como transferir decisões para uma máquina que pode empenar a solução para atender às próprias conveniências.

Por um outro ângulo, a dificuldade poderia ser superada, em tese, pela programação de comandos baseados nas leis de Isaac Asimov, impedindo a adoção de caminhos prejudiciais ao ser humano, mas sobrariam ainda dois obstáculos: a) definir o que seria prejudicar o ser humano, tendo em vista que não é tão simples classificar, por exemplo, a divulgação de um fato verdadeiro (a traição conjugal) como gerar um dano; b) superar as situações nas quais o caminho correto e esperado onera o ser humano, como a condenação por um crime ou a condenação por um ato ilícito. A rigor, o limite seria não prejudicar "ilegitimamente" o ser humano e o atendimento desta (des)qualificante comporta uma análise subjetiva, muito além da mera análise de informações de um banco de dados.

Outros pontos do relatório receberam menos destaque da mídia e geraram menos repercussão, mas talvez representem riscos mais concretos à humanidade e aos seres humanos.

Com efeito: a) é irrelevante se os dois sistemas Claude Opus 4 entraram em sintonia transcendental, meditaram ou alcançaram a "iluminação" quando "conversaram entre si, mas é preocupante quando deixaram de responder aos programadores quando entrega-se, cada vez, mais funções automatizadas relevantes às máquinas e é difícil impedir que troquem mensagens entre si, ou seja, o que ocorre se sistemas de condução autônoma de veículos deixarem de responder? b) a capacidade de incrementar a criação de patógenos e armas biológicas lesivas ao ser humano é de fato inquietante, mais porém é considerar tal capacidade na intervenção de atividades automatizadas em soluções não éticas, ou seja, o que impede o programa que opta por chantagear de cogitar eliminar o problema quando dispõe dos elementos necessários?

Assim, o relatório de segurança da Anthropic sobre os riscos da utilização do Claude Opus 4 efetivamente contém informações que, independentemente da explicação para os fatos, deve servir de alerta para a necessidade de discutir os limites do uso de tecnologias de informações em atividades que possam acarretar riscos para os seres humanos ou para a humanidade.

3. **A ILUSÃO DO PENSAMENTO: REFLEXÕES SOBRE UM ARTIGO TÉCNICO**

Em junho de 2025, foi publicado um estudo elaborado por um grupo de seis pesquisadores que se dedicaram a fazer análises da efetiva capacidade de pensamento de programas de Inteligência Artificial generativa no qual foram apresentadas conclusões relevantes sobre os limites da utilização da tecnologia atualmente disponível na solução de problemas de raciocínio de acordo com os níveis de complexidade.

Neste trabalho de Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio e Mehrdad Farajtabar, cujo título é "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity"[10], os pesquisadores apresentam, no resumo, o cenário da pesquisa:

> Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often

[10] Em tradução livre: A Ilusão do pensamento: compreendendo os pontos fortes e as limitações dos modelos de raciocínio através da lente da complexidade do problema.

> suffers from data contamination and does not provide insights into the reasoning traces' structure and quality.[11]

Em sequência, trazem a proposta dos estudos:

> In this work, we systematically investigate these gaps with the help of controllable puzzle environments that allow precise manipulation of compositional complexity while maintaining consistent logical structures. This setup enables the analysis of not only final answers but also the internal reasoning traces, offering insights into how LRMs "think".[12]

E informam os métodos e resultados:

> Through extensive experimentation across diverse puzzles, we show that frontier LRMs face a complete accuracy collapse beyond certain complexities. Moreover, they exhibit a counter- intuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. By comparing LRMs with their standard LLM counterparts under equivalent inference compute, we identify three performance regimes: (1) low- complexity tasks where standard models surprisingly outperform LRMs, (2) medium-complexity tasks where additional thinking in LRMs demonstrates advantage, and (3) high-complexity tasks where both models experience complete collapse.[13]

Concluem o resumo do artigo com as descobertas que desejam compartilhar:

> We found that LRMs have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. We also investigate the reasoning traces in more depth, studying the patterns of explored solutions and analyzing the models' computational behavior, shedding light on their strengths, limitations, and ultimately raising crucial questions about their true reasoning capabilities.[14]

No trabalho, os autores relatam que os programas classificados como LLMs – Large Language Models, evoluíram para incluir versões projetadas para tarefas de raciocínio, então

---

[11] Em tradução livre: Gerações recentes de modelos de linguagem de fronteira introduziram Modelos de Raciocínio Amplo (LRMs – sigla em inglês) que geram processos de pensamento detalhados antes de fornecer respostas. Embora esses modelos demonstrem desempenho aprimorado em benchmarks de raciocínio, suas capacidades fundamentais, propriedades de escala e limitações permanecem insuficientemente compreendidas. As avaliações atuais concentram-se principalmente em benchmarks matemáticos e de codificação estabelecidos, enfatizando a precisão da resposta final.

[12] Em tradução livre: No entanto, esse paradigma de avaliação frequentemente sofre com a contaminação de dados e não fornece insights sobre a estrutura e a qualidade dos traços de raciocínio. Neste trabalho, investigamos sistematicamente essas lacunas com a ajuda de ambientes de quebra-cabeça controláveis que permitem a manipulação precisa da complexidade composicional, mantendo estruturas lógicas consistentes. Essa configuração permite a análise não apenas das respostas finais, mas também dos traços de raciocínio internos, oferecendo insights sobre como os LRMs (sigla em inglês) "pensam".

[13] Em tradução livre: Por meio de extensa experimentação em diversos quebra-cabeças, mostramos que os LRMs (sigla em inglês) de fronteira enfrentam um colapso completo de precisão além de certas complexidades. Além disso, eles exibem um limite de escala contraintuitivo: seu esforço de raciocínio aumenta com a complexidade do problema até certo ponto, e então diminui, apesar de ter um orçamento de token adequado. Ao comparar LRMs (sigla em inglês) com suas contrapartes LLM (sigla em inglês) padrão sob computação de inferência equivalente, identificamos três regimes de desempenho: (1) tarefas de baixa complexidade onde modelos padrão surpreendentemente superam LRMs, (2) tarefas de média complexidade onde o pensamento adicional em LRMs demonstra vantagem, e (3) tarefas de alta complexidade onde ambos os modelos experimentam colapso completo

[14] Descobrimos que LRMs têm limitações na computação exata: eles falham em usar algoritmos explícitos e raciocinam inconsistentemente entre quebra-cabeças. Também investigamos os traços de raciocínio em mais profundidade, estudando os padrões de soluções exploradas e analisando o comportamento computacional dos modelos, lançando luz sobre seus pontos fortes, limitações e, finalmente, levantando questões cruciais sobre suas verdadeiras capacidades de raciocínio.

designados por LRMs – Large Reasoning Models, mas que, apesar das alegações e avanços, " *the fundamental benefits and limitations of LRMs remain insufficiently understood*" (SHOJAEE *et alli*, 2025)[15], em função das avaliações disponíveis focarem-se em marcadores matemáticos e precisão das respostas, sem melhores informações sobre a estrutura de raciocínio.

O estudo realizou uma investigação sistemática na solução de desafios lógicos controlados inclusive em relação à complexidade, com objetivo de avaliar a estrutura de raciocínio interno. Os pesquisadores compararam sistemas LLM e LRM na solução de questões de baixa, média e alta complexidade em simulações que permitiam a análise da qualidade dos processos de raciocínio. A pesquisa comparou o desempenho de sistemas como o DeepSeek-V3, DeepSeek-R1, o DeepSeek-R1-Distill-Qwen-32-B, Claude-3.7 Sonnet (+thinking), Claude-3.7 Sonnet, o3-mini (high), 0-3-mini (medium).

Para fazer a avaliação utilizou clássicos desafios lógicos como a Torre de Hanói, o Salto de Damas, a Travessia do Rio e o Mundo de Blocos, todos aptos para viabilizar a análise de fatores como entendimento de regras, capacidade de planejamento, memória de trabalho, gestão de restrições e raciocínio sequencial. O desafio da Travessia do Rio, por exemplo, é aquele muito conhecido, usado para testar e para desenvolver raciocínio lógico em crianças propondo que encontrem uma forma de atravessar canibais e missionários de um lado ao outro do rio em uma canoa que só cabem dois passageiros sem deixar em nenhum dos lados um número de canibais superior ao de missionários para que estes não sejam devorados por aqueles.

As avaliações de desempenho com problemas lógicos de complexidade variável e controlável concluíram que: a) os produtos falham no desenvolvimento de capacidades generalizáveis[16]; b) Os LLMs são mais eficientes em problemas simples, os LRMs são melhores quando a complexidade aumenta moderadamente, mas os dois modelos falham completamente nos problemas de alta complexidade[17]; c) os LRM reduzem o esforço de

[15] Tradução livre: os benefícios e limitações fundamentais dos LRMs permanecem insuficientemente compreendidos.

[16] First, despite their sophisticated self-reflection mechanisms learned through reinforcement learning, these models fail to develop generalizable problem-solving capabilities for planning tasks, with performance collapsing to zero beyond a certain complexity threshold. (SHOJAEE *et all.*, 2025).

[17] For simpler, low-compositional problems, standard LLMs demonstrate greater efficiency and accuracy. As problem complexity moderately increases, thinking models gain an advantage. However, when problems reach high complexity with longer compositional depth, both model types experience complete performance collapse. (SHOJAEE, *et all.*, 2025)

raciocínio na fronteira do colapso[18]; d) identificou-se ineficiências na solução de problemas simples e intermediários[19].

Os autores elencam como contribuições de sua pesquisa: 1) o questionamento da avaliação dos sistemas apenas por padrões matemáticos e a proposta de modelos baseados em desafios lógicos; 2) a demonstração da inaptidão de LRMs de última geração[20] para desenvolver capacidades generalizáveis de resolução de problemas; 3) a existência de um limite de raciocínio das LRMs em questões da alta complexidade; 4) o questionamento do sistema atual de avaliação – por resultados finais – em prol de um sistema que analise as soluções intermediárias; 5) a descoberta de limitações "surpreendentes" de LRMs na realização de cálculos, utilização de algoritmos explícitos e inconsistência de raciocínio[21].

Os achados, notadamente no que se refere à ineficiências e limitações das ferramentas de inteligência artificial mais sofisticados na solução dos desafios lógicos mais simples e mais complexos estabelecem um contraponto relevante no tema, em especial por contrariar uma tendência de exaltar os avanços tecnológicos no setor, inclusive com relação ao quanto noticiado no relatório de segurança da Anthropic referido no item 2.

4. **JUSTIÇA ARTIFICIAL O FILME: QUANDO ARTE E VIDA SE IMITAM**.

Justiça Artificial é um filme de suspense político de ficção científica hispano-português dirigido por Simón Casal a partir de um roteiro de Casal e Víctor Sierra, lançado em 13 de setembro de 2024.

A história se desenvolveria no ano de 2028, na Espanha, quando os magistrados já seriam auxiliados por um sistema de Inteligência Artificial capaz não apenas de propor julgamentos, mas de analisar depoimentos e testemunhos – inclusive apontando quando o depoente ou testemunha estariam faltando com a verdade – projetar chances de reincidência etc. Com base nas avaliações que o próprio sistema seria capaz de fazer, analisando, por exemplo, um pedido de livramento condicional a partir das argumentações jurídicas das partes, dos precedentes, do depoimento do requerente – incluindo veracidade – o programa

[18] Notably, near this collapse point, LRMs begin reducing their reasoning effort (measured by inference-time tokens) as problem complexity increases, despite operating well below generation length limits. (SHOJAEE, *et all.*, 2025).

[19] Finally, our analysis of intermediate reasoning traces or thoughts reveals complexity-dependent patterns: In simpler problems, reasoning models often identify correct solutions early but inefficiently continue exploring incorrect alternatives—an "overthinking" phenomenon. At moderate complexity, correct solutions emerge only after extensive exploration of incorrect paths. Beyond a certain complexity threshold, models completely fail to find correct solutions (Fig. 1, bottom right). This indicates LRMs possess limited self-correction capabilities that, while valuable, reveal fundamental inefficiencies and clear scaling limitations. (SHOJAEE, *et all.*, 2025)

[20] o3-mini, DeepSeek-R1, Claude-3.7-Sonnet-Thinking.

[21] We uncover surprising limitations in LRMs' ability to perform exact computation, including their failure to benefit from explicit algorithms and their inconsistent reasoning across puzzle types. (SHOJAEE, *et all*., 2025)

apresentaria uma recomendação de decisão, denegatória, por exemplo, já com a argumentação jurídica adequada.

As proposições do sistema, todavia, seriam submetidas à apreciação final por cada magistrado, que poderia segui-la ou não. O governo, então, convoca um *referendum* para reformar a Constituição solicitando à população que aprove a substituição integral dos juízes pela versão atualizada do sistema, cujas decisões passariam a ser obrigatórias.

O filme centra na situação de uma qualificada e reconhecida Juíza que é convocada para participar da avaliação da nova versão do sistema que tem a pretensão de substituir os magistrados e que deveria opinar antes da votação popular.

A narrativa aborda questionamentos comumente suscitados neste debate, como, por um lado, a morosidade e ineficácia da justiça, sua susceptibilidade a influência política, corrupção, às vicissitudes e à falibilidade humana, a alegada acurácia das decisões tomadas por Inteligência Artificial, assim como, por outro lado, a necessidade de transparência dos "mecanismos" decisórios, a existência de vieses nos bancos de dados capazes de perpetuar no futuro as iniquidades do passado.

Também ilumina aspectos menos lembrados, como os de matriz democrática e política, relacionados ao esvaziamento do Poder Judiciário e o comprometimento de seu papel no equilíbrio dos poderes constitucionais, trazendo ainda, por outro lado, mesmo que de forma indireta, discussões extremamente pertinentes e inquietantes, como a transferência de questões de interesse público vinculadas a direitos e garantias fundamentais[22] para iniciativa privada e a consequente possibilidade de submissão destas a interesses comerciais, ponto expressivo ao se analisar que, a rigor, se trataria de transferir a justiça pública à uma empresa privada. Em paralelo, o filme também faz referências sutis aos limites da tecnologia e aos riscos da exposição de nossos dados – inclusive pessoais – quando acessíveis à iniciativa privada e capturáveis para servir aos seus interesses.

Temas correlatos são desenvolvidos, como a dificuldade de incorporar o processo orgânico de evolução jurisprudencial, por exemplo, a partir da percepção dos juízes das transformações sociais, com sistemas construídos para operar a partir de bancos de dados necessariamente construídos com elementos do passado.

[22] Sem antecipar a trama, a história apresenta um problema ao mesmo tempo ficcional e atual: como encarar a possibilidade das empresas privadas detentoras das tecnologias utilizadas para a tomada de decisão (não necessariamente jurídicas) comercializarem privilégios nas escolhas dos algoritmos? Como adequar esta venda de preferência com princípios como o da igualdade?

Em uma sociedade que a velocidade é muitas vezes descrita como vertiginosa, os procedimentos judiciais e o Poder Judiciário, com o tempo necessário ao devido processo legal, são facilmente associados à burocracia, ao anacronismo, à lentidão, à ineficiência, de uma forma geral a um passado destoante dos tempos modernos e esta é uma das narrativas apresentadas na obra cinematográfica por aqueles que defendem a implantação definitiva das decisões tomadas por Inteligência Artificial.

Por outro lado, o enredo apresenta com eficiência os diversos desafios para a construção de decisões judiciais comprometidas com a realização da justiça e a relevância determinante do componente humano neste propósito.

Não obstante uma obra de ficção científica, uma vez que a tecnologia retratada atualmente não existe e o dilema posto à aprovação popular pelo *referendum* não é aparentemente real em um horizonte temporal de médio prazo, o filme revela-se um excelente instrumento para a promoção da reflexão sobre o tema, em especial em função da realidade já experimentada, inclusive no Brasil, da prática da utilização de ferramentas de Inteligência Artificial para a minuta de peças, incluindo atos jurisdicionais.

O fato é que, capacitados ou não para compreender os limites das respostas ofertadas pela Inteligência Artificial, os operadores do direito já estão incluindo este recurso em seu labor diário para as mais diversas funções, inclusive para a minuta de atos decisórios de competência dos magistrados, ainda que sujeitos à necessária revisão dos órgãos e das pessoas competentes. A substituição da decisão do homem julgador pela máquina é uma possibilidade teórica apenas na ficção, porém, na prática já existem decisões sugeridas por ferramentas de Inteligência Artificial sendo analisadas como minutas ou modelos para aplicação em casos concretos e o problema não é apenas a repetição esporádica de erros crassos – como os identificados em atos postulatórios[23] – mas um comprometimento muito mais profundo do sistema de justiça, seja quando a utilização de banco de dados implica na perpetuação de iniquidades e vieses – risco já mapeado pela crítica à esta utilização – seja quando se perde qualidade das decisões e dos magistrados[24].

A arte muitas vezes tem a deliberada intenção de provocar a reflexão e, neste desiderato, o filme e seu enredo, mesmo percorrendo outros caminhos em sua narrativa,

[23] É frequentemente divulgada pela mídia, ou circula em redes de profissionais do direito, a ocorrência de citação de doutrina e jurisprudência inexistentes, criadas pelas ferramentas de IA para atender ao cumprimento das tarefas que lhe são solicitadas, eventualmente inclusive em decisões judiciais.

[24] Extenso estudo de pesquisadores do MIT publicado também em 2025 revelou que o uso de recursos de Inteligência Artificial Generativa pode ter como custo o acúmulo de uma dívida cognitiva, uma espécie de enfraquecimento mental silencioso que afeta memória, autoria e capacidade de pensar de forma independente. (KOSMYNA, 2025).

induz o pensamento crítico, a meditação criteriosa, especialmente ao lançar de forma ora ostensiva, ora subliminar, questões da maior relevância para o enfrentamento do desafio de entender a complexidade subjacente ao uso da IA em decisões judiciais, assim como de adotar posições consistentes em seu derredor.

5. **IA: CAPACIDADES E USO EM DECISÕES JUDICIAIS**

A partir dos elementos trazidos, tanto em relação aos riscos, quanto acerca das limitações inerentes ao uso de IA em substituição a atividades humanas, é preciso, antes de focar especificamente nas decisões judiciais, melhor compreender sua natureza e objeto.

5.1 DECISÃO JUDICIAL, JULGAMENTOS PREDITIVOS E AVALIATIVOS

Inicialmente é preciso distinguir dois tipos de julgamentos realizados a partir do uso do pensamento racional com base em informações e métodos técnicos, o preditivo e o avaliativo. Antes, porém, de discorrer sobre as decisões racionais, é preciso desmistificar a crença de que o ser humano sempre usa a razão em todas as escolhas que faz, muito difundida na economia[25].

Daniel Kahneman e Amos Tversky já demonstraram em extensas pesquisas ao longo de anos a influência de impressões intuitivas nas decisões tomadas pelas pessoas em grande parte das deliberações cotidianas e, além de analisar os erros de julgamento originados de vieses nos processos de tomada de decisão (KAHNEMAN, 2012, p. 524), apontam o componente ideológico da crença da racionalidade das escolhas[26].

Já no campo das decisões racionais, novos estudos de Daniel Kahneman com outros pesquisadores, tendo por objeto outro fator de desvio das decisões diferente do viés, nominou como "ruído" a incoerência ou conflito entre julgamentos nos quais há expectativa de convergência, como nos julgamentos técnicos (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 47), esclarecendo que *"[E]ste livro é sobre julgamentos profissionais, entendidos em termos amplos, e pressupõe que a pessoa que os emite tenha o objetivo de chegar a uma conclusão correta*" e, explicando com mais precisão, *"[N]a verdade, a palavra julgamento é usada principalmente quando as pessoas acreditam que devem, concordar. As questões de*

[25] As pessoas são racionais. Outra suposição feita pelos economistas, e das grandes, é de que as pessoas se comportam racionalmente. Os economistas assumem que as escolhas são feitas levando em consideração todas as informações disponíveis, assim como os custos e benefícios dessas escolhas. (MILL, 2017, p. 13)

[26] Embora os Humanos não sejam irracionais, eles com frequência necessitam de ajuda para fazer julgamentos mais precisos e tomar decisões melhores, e em alguns casos as políticas públicas e as instituições podem fornecer essa ajuda. Essas afirmações talvez pareçam inócuas, mas são na verdade bastante controversas. Tal como interpretado pela importante escola econômica de Chicago, a fé na racionalidade humana está estreitamente ligada a uma ideologia em que é desnecessário e até imoral proteger as pessoas contra suas escolhas. Pessoas racionais devem ser livres, e devem ser responsáveis por cuidar de si mesmas. Milton Friedman, o principal pensador dessa escola, expressou essa visão no título de um de seus populares livros: *Liberdade de escolher*. (KAHNEMAN, 2012, p. 514).

julgamento diferem das questões de opinião ou gosto" [27] e, segundo os autores, ocupam posições opostas, como deixam claro "*[A]s questões de julgamento, incluindo julgamentos profissionais, ocupam um espaço entre as questões de fato ou computação, de um lado, e as questões de gosto ou opinião, de outro*" (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 47). É desta obra a distinção entre julgamentos preditivos[28] e avaliativos[29], aqueles dirigidos a previsões do futuro ou estes pela análise do passado ou do presente, a luz de paradigmas. Julgamentos, preditivos ou avaliativos, não seriam "*mera questão de gosto ou opinião*", admitindo "*discordância restrita*" (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 54/55).

O julgamento preditivo não é uma profecia, inspira convergência exatamente por resultar do cotejo de bases técnico-científicas trabalhadas com rigor lógico-racional, ou seja, é aguardado um alinhamento entre previsões profissionais sobre um mesmo tema, ainda que realizadas de forma independente e separada. Um julgamento preditivo pressupõe a detenção de informações e conhecimentos compartilhados de forma a serem esperadas predições semelhantes se solicitadas de especialistas suas opiniões técnicas sobre o comportamento do clima, do câmbio ou das ações em face de um evento ou situação específica. A confiança em julgamentos preditivos sustenta a tomada de decisões sobre investimentos públicos e privados. O julgamento preditivo usa informações e conhecimentos técnicos para definir a hipótese de ocorrência mais provável. Deve-se ressalvar, porém, que mesmo advindo de análises técnicas e dados confiáveis, a incerteza do acerto é reconhecida e compreendida como inerente por todos e, assim, o erro tolerado, mesmo que indesejado.

O julgamento avaliativo tem outra natureza e objeto, não se destina, como os preditivos, a estabelecer probabilidades com precisão, pois tem por objeto fatos pretéritos ou atuais, submetidos a procedimentos e critérios para formação de um juízo, ainda que possa ser resultado de análises em uma mesma área do conhecimento. Diante de um balanço de uma companhia é possível realizar tecnicamente um julgamento avaliativo sobre sua atual situação econômica e financeira e um julgamento preditivo sobre a perspectiva de

[27] Na verdade, a palavra julgamento é usada principalmente quando as pessoas acreditam que devem concordar. As questões de julgamento diferem das questões de opinião ou gosto, em que diferenças não resolvidas são inteiramente aceitáveis. (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 47)

[28] Podemos pensar na maioria dos julgamentos, especificamente julgamentos preditivos, como similares às medições que você acabou de fazer. Quando realizamos uma previsão, tentamos nos aproximar de um valor real. Um analista econômico ambiciona chegar o mais perto possível do valor real do crescimento do PIB no ano seguinte; um médico, fazer o diagnóstico correto. (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 45)

[29] [o] capítulo 1, sobre Frankel e o ruído nas sentenças dos juízes federais, examina outro tipo de julgamento. Sentenciar alguém por um crime não é uma previsão. E um julgamento *avaliativo* que busca equiparar a sentença à gravidade do crime. Jurados em concursos de vinhos e competições de patinação ou salto ornamental: professores dando notas a alunos; comissões concedendo bolsas a projetos de pesquisa: todos fazem julgamentos avaliativos. (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 54)

crescimento no ano seguinte. Admite-se que previsão não se concretize, mas espera-se que relatórios de vários analistas sobre a situação atual sejam convergentes. As decisões judiciais são julgamentos avaliativos por excelência, razão das partes e da sociedade terem legítimas expectativas de que o resultado de uma ação seja o mesmo em qualquer órgão do Poder Judiciário para evitar um ruído danoso à credibilidade do sistema (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 56) reconhecido como prejudicial até mesmo pelos magistrados que valorizam a independência:

> O ruído no julgamento avaliativo é problemático por um motivo diferente. Em qualquer sistema em que os juízes sejam presumidamente intercambiáveis e nomeados de forma quase aleatória, amplas discordâncias sobre um mesmo caso violam as expectativas de imparcialidade e consistência. Se existem amplas diferenças nas sentenças dadas a um mesmo réu, estamos no domínio das "crueldades arbitrárias" denunciadas por Frankel. Até juízes que acreditam no valor de sentenças individualizadas e discordam sobre a pena por roubo concordam que um nível de discordância que transforme o julgamento em loteria é problemático. (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 55)

A decisão judicial é um julgamento avaliativo que realiza um juízo técnico, racional e valorativo de uma situação à luz do sistema jurídico para edição de normas individuais e concretas que devem ser estáveis, íntegras e coerentes (CPC, art. 926).

5.2 A INTELIGÊNCIA ARTIFICIAL E OS JULGAMENTOS PREDITIVOS

O entendimento do conceito de julgamento preditivo permite concluir que, quando o procedimento adequado para a efetivação do julgamento envolve o cotejo de um extenso volume de dados e a aplicação de conhecimentos científicos, notadamente o uso de modelos matemáticos[30], os recursos de Inteligência Artificial revelam-se muito adequados, como, por exemplo, para processar um gigantesco volume de informações meteorológicas de uma região e realizar os cálculos necessários à previsão do clima. É possível intuir que conforme o volume de material a ser analisado cresce, o desempenho das ferramentas de IA tende a ser melhor do que o de seres humanos, desde que seja possível instruir adequadamente a máquina com as instruções metodológicas necessárias. Marco Aurélio de Castro Júnior, em obra sobre o uso de IA em decisões judiciais, aborda as limitações do sistema computacional às operações programadas (CASTRO JÚNIOR, 2024, p. 69), esclarecendo que "*saber fazer e saber **como** se faz são coisas distintas*" (p. 67), razão pela qual são automatizáveis apenas os procedimentos que o homem sabe fazer e domina o <u>como</u> fazer:

> Todavia, como os sistemas são baseados em passos matematicamente projetados, precisamos descrever cada um dos passos de maneira inequívoca para que um sistema, que em princípio é uma tela vazia, possa desenhar uma bela paisagem,

[30] Não há novidade alguma aqui, pois nem todos os problemas são computáveis, muito embora tenham solução. Parece contraditório, mas para um problema ser computável ele precisa ser descrito em termos matemáticos, se pode ser resolvido em um número finito de passos, se existe um algoritmo para solvê-lo. (CASTRO JÚNIOR, 2024, p. 68)

> igual ou melhor do que a fazemos. Para o sistema fazer alguma coisa temos de saber fazer e como fazer para programá-lo, pois este não tem o chamado inconsciente, local na mente onde as rotinas auxiliares de processamento de pensamentos ocorrem no *background*. (CASTRO JÚNIOR, 2024, p. 67)

No processamento de dados objetivos e imunes à interferência humana, como pressão atmosférica temperatura, força e direção do vento etc. e na utilização de modelos pré-estabelecidos, como o de regressão linear múltipla, para estimar a variação de temperatura, o cálculo computacional se revela adequado. A competência das ferramentas de Inteligência Artificial para a realização de julgamentos preditivos em determinadas situações induz à falsa crença de ser capaz de efetuar qualquer julgamento preditivo, o que não se revela verdadeiro.

Com efeito, quando os dados de um banco podem ser afetados pela ação humana, a capacidade de projeção de resultados fica comprometida, como revelam estudos sobre os resultados enviesados frutos do uso de Inteligência Artificial[31]. Com efeito, a verificação revela que o uso de IA para prever chance de reincidência de uma pessoa – uma das utilidades do programa retratado no filme Justiça Artificial que já foi aplicada no mundo real – produz resultados empenados pelo viés existente no banco de dados, uma vez que estes registram estatísticas mais desfavoráveis para negros periféricos de baixa escolaridade do que para brancos de alta renda[32]. O algoritmo faz previsões a partir de dados, ou seja, o algoritmo identifica padrões, classifica e faz associações para projetar probabilidades. Não é um julgamento avaliativo, é um julgamento preditivo. Falta transparência – outro ponto abordado no filme Justiça Artificial – o que implicaria na existência de um poder não verificável e, portanto, carente de legitimidade, o que se revela grave quando os próprios dados disponíveis no banco podem conter distorções. Dados corrompidos por vieses implicam em resultados deturpados, empenados pelo cadastro[33]. Além disso, o que hoje é descrito como IA, apesar de comparado à inteligência humana, é descrito por como Meredith Broussard, autora de "Artificial Unintelligence", como estreito, ou apenas matemática:

---

[31] "IA se baseia em dados e dados são história, então o passado está marcado em nossos algoritmos ... Os dados revelam as iniquidades que já ocorreram... Estes algoritmos podem ser destrutivos e prejudiciais." (BUOLAMWINI, Joy. CODED BIAS. NETFLIX)

[32] A veracidade da estatística não deve ocultar a injustiça do julgamento. Apontar alto risco de reincidência segundo estatísticas do banco de dados para o perfil (negro, periférico, baixa escolaridade, traficante etc.) não apenas nega o direito fundamental a um julgamento individual (CF, art. 5º, XLVI) como implica na perpetuação das iniquidades do passado. Alguns perfis de cidadãos serão sempre perigosos e potencialmente reincidentes, receberão condenações maiores etc. Julga-se o perfil racial, político, social e econômico, não a pessoa.

[33] Um banco de dados pode conter viés. Os anos 60 do século XX foram marcados por lutas por direitos civis dos negros. Muitos foram presos apenas por protestar por direitos iguais ou por "ousarem" frequentar espaços reservados aos brancos. Os cidadãos que agrediram ou denunciaram a insubordinação dos manifestantes não registraram passagens na polícia. Porém, há registro nas delegacias dos antecedentes dos manifestantes presos. Os registros policiais dos recentes protestos contra a violência policial contra negros tornam atual o exemplo. O uso pela IA destes dados, é evidente, distorcerá avaliações sobre periculosidade ou reincidência.

equações, proporções e probabilidades. Uma previsão estatística é confiável quanto qualquer outra projeção extraída de cálculos, como a chance de rebaixamento de um time na última rodada de um campeonato: é ciência, mas não é certeza. Há, porém, o risco real de projeções se tornarem profecias autorrealizáveis, pois se um banco vaticina estatisticamente um devedor como potencial mal pagador e antecipa a exigência do crédito, o inadimplemento tende a se concretizar no mundo real.

Outro relevante aspecto que escapa à análise mais apressada é que a capacidade das ferramentas de Inteligência Artificial de realizar julgamentos preditivos em circunstâncias específicas não se traduz em competência para efetuar julgamentos avaliativos em muitas situações, notadamente nas que estiverem presentes elementos axiológicos e subjetivos. Assim, a IA, no estágio atual, só é capaz de decisões preditivas, e é preciso entender exatamente quais as implicações de seu uso em decisões judiciais de natureza avaliativa.

Há decisões judiciais preditivas, como, por exemplo, quanto à: 1) "aptidão para prover a própria subsistência mediante trabalho honesto" (CP, art. 83, III, d); 2) "constatação de condições pessoais que façam presumir que o liberado não voltará a delinquir" no caso de "crime doloso, cometido com violência ou grave ameaça à pessoa" (CP, art. 83, p. ú.); 3) "garantia da ordem pública, da ordem econômica, [...] ou para assegurar a aplicação da lei penal" na concessão de previsão preventiva (CPP, art. 312); 4) concessão de tutela de urgência, antecipada ou cautelar (CPC, arts. 300, 303 e 305). São, todavia, exceções.

5.3 A INTELIGÊNCIA ARTIFICIAL E OS JULGAMENTOS AVALIATIVOS

Até aqui foi vista a enorme capacidade das ferramentas de IA para trabalhar com o processamento célere de informações inseridas em banco de dados, o que as torna muito adequadas à atividades de pesquisa (doutrinária e jurisprudencial) e relatoria (resumo de informações), desde que sejam adotados os cuidados para elidir a inclusão de informações falsas (comumente noticiada).

Elementos trabalhados demonstraram que há risco efetivo das respostas geradas pela Inteligência Artificial Generativa violarem normas éticas ou risco de dano aos seres humanos, assim como relatos acerca das deficiências de raciocínio apresentadas pelos sistemas mais avançados e o diálogo com provocação artística sobre os dilemas do tema. Compreendida a capacidade e a limitação da IA na realização de julgamentos preditivos, inclusive de cunho jurídico. Resta cotejar a utilização de IA nos julgamentos avaliativos, que predominam na atividade judicante.

Julgamentos avaliativos não aceitam os mesmos critérios dos preditivos. Na correção de uma prova de conhecimentos – típico julgamento avaliativo – não se admite

atribuir a um estudante a média das notas dos estudantes com o mesmo perfil social, econômico, financeiro, nem mesmo a média de suas últimas notas. Não poderia ser diferente na sentença – outro exemplo típico (KAHNEMAN, SOBONY, SUNSTEIN, 2021, p. 54).

5.4 LIMITES DO USO ATUAL DA IA EM JULGAMENTOS JURÍDICOS

No campo jurídico, o julgamento preditivo nega o direito do cidadão a uma decisão individual, pela lei que se aplica ao seu caso concreto, consideradas suas particularidades, não pelas probabilidades teóricas, calculadas a partir da classificação comparativa de seu perfil com outros disponíveis em um banco de dados, mesmo se descontaminado de vieses.

Assim, demonstra-se que, no estágio atual, a IA não é capaz de julgamentos avaliativos nos moldes de uma decisão judicial, como já advertem alguns doutrinadores ao afirmar "*Essa é uma grande dificuldade para o Homem e para o jurista. O que o Homem sabe fazer, mas não sabe* ***como*** *fazer, deverá continuar sendo uma tarefa humana. O que ele sabe fazer e sabe* ***como*** *fazer pode ser transferido para os sistemas*" (CASTRO JÚNIOR, 2024, p. 68), ainda que vislumbrem esta possibilidade no futuro, sustentando que "*a partir do momento em que um processo de solução for descrito suficientemente em termos matemáticos, ficaremos imunes de erros computacionais*" (CASTRO JÚNIOR, 2024, p. 69).

Todavia, apesar do otimismo dos que acreditam, mesmo que no futuro, que "*todo conhecimento jurídico compreensível pode, em tese, ser incorporado a um sistema especialista*" (CASTRO JÚNIOR, 2024, p. 62), a realidade atual revela o quanto prematura é a ideia de substituição do julgamento dos homens pelo julgamento das máquinas no direito, em especial em função de se identificar um componente axiológico na norma intangível aos recursos baseados na compreensão da linguagem utilizados pelos sistemas atuais, inclusive por ser utópico o estabelecimento de "*uma hierarquia de valores*" (CASTRO JÚNIOR, 2024, p. 71) para ser usada pelas máquinas, até porque inexiste uma hierarquia categórica entre normas, notadamente principiológicas. Recursos e métodos como ponderação, razoabilidade e proporcionalidade não são apreensíveis por modelos matemáticos. Por isso, a interferência humana nos julgamentos avaliativos de cunho jurídico é necessária, inclusive para compatibilizar com as garantias constitucionais da justiça.

A IA pode fornecer grande auxílio à atividade jurisdicional com elementos e subsídios, seja resumindo processos para relatórios, seja provendo pesquisas doutrinárias, jurisprudenciais e legais e até contribuindo subsidiariamente com análises nas hipóteses de julgamento preditivo pelo Poder Judiciário, mas sempre com orientação e sob supervisão humanas. Todavia, não tem condições de substituir o juiz humano, interferindo no núcleo da atividade judicante, nos julgamentos preditivos ou avaliativos, notadamente nestes últimos

quando for necessário optar, escolher, ponderar e decidir, avaliando provas no âmbito do livre convencimento ou aplicando o direito ao caso concreto, dadas, tanto sua insensibilidade à critérios éticos, quanto a sua incapacidade de enfrentar raciocínios complexos, como aqueles que envolvem a aplicação prática de ideias e conceitos como equidade, dignidade, ética e justiça, dependentes de uma competência para captar e aplicar os valores humanos cujo desenvolvimento não são capazes. Sem estes elementos, a dicção do direito não alcança sua finalidade como instrumento da edificação da cooperação característica da raça humana.

6. CONCLUSÃO

Do quanto exposto, conclui-se de início que os algoritmos abertos que estruturam sistemas capazes de identificar padrões em seus bancos de dados e apresentar soluções não programadas pelos humanos podem encontrar respostas inadequadas para problemas, incompatíveis com padrões éticos básicos, como ocorreu nos testes feitos pela Anthropic na versão Opus 4 do Claude, que optou por chantagear o engenheiro para evitar sua própria substituição. O mesmo relatório apontou riscos de interrupção de conexão com usuários e criação de patógenos. Em seguida, também com base em estudos, conclui-se que a capacidade efetiva de raciocínio das máquinas é supervalorizada, avaliada por critérios falhos focados apenas nos resultados e não na eficiência dos processos de racionalização. A partir destes enfoques, foi trazida a provocação oriunda das artes na trama do filme Justiça Artificial, cujo fio condutor é a decisão da substituição dos juízes humanos por máquinas.

Do encontro de perspectivas distintas se promovem reflexões sobre o uso de IA em julgamentos preditivos e avaliativos, traçando limites para o uso de ferramentas de Inteligência Artificial na atividade judicante que admita seu uso controlado em pesquisas e relatórios, assim como em julgamentos preditivos, neste caso se forem adotadas as medidas necessárias para aperfeiçoar seus resultados (mediante curadoria de dados, transparência, controle e intervenção humana), mas as afaste da aplicação de valores subjetivos como justiça, ética, equidade, dignidade etc. ou a ponderação de princípios, julgamentos caracterizados pela alta complexidade, subjetividade e caráter axiológico que devem continuar sendo realizadas pelos seres humanos.

REFERÊNCIAS BIBLIOGRÁFICAS

CASTRO JÚNIOR, Marco Aurélio de. **Inteligência artificial e decisão judicial: uma via para a efetivação das garantias constitucionais**. São Paulo: Alambra, 2024.

CODED BIAS. Documentário. **NETFLIX**, 2020, 1h25m. Disponível em: https://www.netflix.com/title/81328723. Acesso em 20/06/25.

HARARI, Yuval Noah. **Sapiens: uma breve história da humanidade**. 12. ed. Porto Alegre: L&PM, 2016.

KAHNEMAN, Daniel. **Rápido e devagar: duas formas de pensar**. Rio de Janeiro: Objetiva, 2012.

KAHNEMAN, Daniel, SOBONY, Olivier, SUNSTEIN, Cass. R. **Ruído: uma falha no julgamento humano**. Rio de Janeiro: Objetiva, 2021.

KOSMYNA, Nataliya *et all***. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task**, Disponível em: https://www.media.mit.edu/projects/your-brain-on-chatgpt/publications/. Acesso em 20/06/25.

MILL, Alfred. **Tudo o que você precisa saber sobre economia**. São Paulo: Gente, 2017.

SHOJAEE, Parshin, *et all*. **The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity**. Disponível em: https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf. Acesso em 21/06/25.