UNIVERSIDADE DE BRASÍLIA
FACULDADE DE DIREITO
PROGRAMA DE PÓS-GRADUAÇÃO EM DIREITO

SUZIANY VENANCIO DO ROSARIO

O USO DE INTELIGÊNCIA ARTIFICIAL NO PODER JUDICIÁRIO BRASILEIRO E OS ASPECTOS ÉTICOS:

Uma Análise do Projeto de Pesquisa e Desenvolvimento Sabiá

BRASÍLIA

2024

SUZIANY VENANCIO DO ROSARIO

O USO DE INTELIGÊNCIA ARTIFICIAL NO PODER JUDICIÁRIO BRASILEIRO E OS ASPECTOS ÉTICOS:

Uma Análise do Projeto de Pesquisa e Desenvolvimento Sabiá

Dissertação apresentada ao Programa de Pós-Graduação em Direito da Universidade de Brasília (UnB) como requisito parcial para a obtenção do título de Mestre em Direito.
Área de concentração: Direito, Estado e Constituição.
Orientador: Prof. Dr. Fabiano Hartmann Peixoto.

BRASÍLIA

2024

SUZIANY VENANCIO DO ROSARIO

O USO DE INTELIGÊNCIA ARTIFICIAL NO PODER JUDICIÁRIO BRASILEIRO E OS ASPECTOS ÉTICOS:

Uma Análise do Projeto de Pesquisa e Desenvolvimento Sabiá

Dissertação apresentada ao Programa de Pós-Graduação em Direito da Universidade de Brasília (UnB) como requisito parcial para a obtenção do título de Mestre em Direito.
Área de concentração: Direito, Estado e Constituição.
Orientador: Prof. Dr. Fabiano Hartmann Peixoto.

Aprovação em: 27/09/2024

BANCA EXAMINADORA:

Professor Doutor Fabiano Harmann Peixoto
(Orientador - Presidente da Banca Examinadora)

Professora Doutora Fernanda de Carvalho Lage
(Membro interno da Banca Examinadora - UnB)

Professora Doutora Cristina Mendes Bertoncini Corrêa
(Membro externo da Banca Examinadora - UFSC)

Professor Doutor Henrique Araújo Costa
(Membro suplente da Banca Examinadora - UnB)

Dedico este trabalho à minha família, pelo amor incondicional.

AGRADECIMENTOS

Cursar o Mestrado na Universidade de Brasília é a realização de um sonho, por isso quero expressar a mais sincera gratidão a todos aqueles que de alguma forma contribuíram para esse enriquecedor processo de aprendizagem.

Agradeço, primeiramente, a Deus pela minha vida, por permitir que eu alcance meus objetivos, pela força diária e por sempre iluminar os meus caminhos nos momentos mais difíceis.

Aos meus pais, Suzane e Anisio, e ao meu irmão Caio, sou grata pelo carinho, compreensão e suporte em todas as situações. Vocês são a minha base e fazem os meus dias mais felizes.

Agradeço ao meu orientador, professor Fabiano Hartmann Peixoto, pelos conselhos, paciência, acolhimento e toda a atenção prestada ao longo dos últimos dois anos. Também sou grata à professora Debora Bonat, pela convivência, ensinamentos e pela oportunidade única de fazer parte da equipe de um projeto de IA.

Quero expressar minha profunda gratidão às professoras Fernanda Lage, Cristina Bertoncini e ao professor Henrique Araújo, por prontamente aceitarem o convite para participar da banca examinadora da minha dissertação e pelo tempo dedicado à leitura e avaliação do meu trabalho.

Agradeço aos amigos que me acompanharam ao longo dessa jornada, aos professores por transmitirem conhecimentos com tanta competência, aos funcionários da Universidade Brasília pela dedicação e aos membros dos Laboratórios de Pesquisa DR.IA e AI.Lab, pelo ambiente de pesquisa de excelência e inovação.

RESUMO

Com um potencial transformador em diversos domínios, a inteligência artificial tem provocado reflexões quanto ao seu papel de centralidade e inovação na área do Direito. Verificado o uso da IA pelo Poder Judiciário, através de sistemas que otimizam e incrementam o fluxo das atividades tendentes à resolução de conflitos, o presente trabalho objetivou analisar os parâmetros éticos e regulações que orientam o desenvolvimento e uso desses sistemas no Brasil. Além de uma compreensão geral da temática, realizou-se o exame específico do Projeto de Pesquisa e Desenvolvimento Sabiá, a fim de verificar empiricamente o recorte proposto em uma ferramenta idealizada pela Universidade de Brasília em parceria com o Tribunal Superior do Trabalho. Para atender aos fins apresentados, o primeiro capítulo é dedicado à apresentação de conceitos básicos na área da IA; no capítulo seguinte foram demonstradas as conexões desse campo com o Direito; o terceiro capítulo enfatizou assuntos relativos à transparência e ética desses sistemas e, no último capítulo, procedeu-se ao exame do Projeto Sabiá. Concluiu-se que na ausência de normas, os princípios éticos assumem um caráter orientador na elaboração dos modelos e que com o esforço conjunto de desenvolvedores, usuários, afetados, governos e academia é possível construir um ambiente favorável à IA ética. Quanto à metodologia, o procedimento utilizado é o dedutivo, com o emprego da abordagem qualitativa e das técnicas de pesquisa bibliográfica e documental.

Palavras-chave: Inteligência artificial e Direito; Pesquisa e Desenvolvimento (P&D); Ética da IA; Projeto Sabiá.

ABSTRACT

Known for having a powerful effect in several areas, artificial intelligence has provoked remarks about its central and innovative role in Law. Once verified the use of AI by Judiciary, through systems that optimize and improve the flow of activities intended to resolve conflicts, this master's thesis aimed to analyze ethical frameworks and regulations that guide the development and use of these systems in Brazil. Beyond a general understanding of the subject, a specific examination of the Sabiá Research and Development Project was done to empirically verify this approach in a tool designed by the University of Brasília in partnership with the brazilian Superior Labor Court. For this purpose, the first chapter is dedicated to present general concepts in the area of AI; the next chapter demonstrates the connections between this field and Law; the third chapter emphasizes issues related to the transparency and ethics of these systems; and, in the last chapter, the Sabiá Project was explained. In conclusion, the absence of specific legislation means that ethical principles assumes a guiding role in the development of models; in addition, the efforts of developers, users, stakeholders, governments and academia make it possible to create an auspicious environment to ethical AI. Regarding the methodology, it was used the deductive procedure, combined with qualitative approach and bibliographic and documentary research techniques.

Keywords: Artificial intelligence and Law; Research and Development (R&D); AI ethics; Sabiá Project.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Associando-se à abertura de caminhos e à superação de limites nas mais diversas áreas, a inteligência artificial ganha cada vez mais notoriedade pelo incremento que propicia em diferentes tarefas. O Direito não ficou de fora dessa realidade e o Poder Judiciário destaca-se pela utilização de dezenas de ferramentas de IA nos tribunais. O surgimento de preocupações acerca da ética dos sistemas acompanha o crescimento desse fenômeno inovador, de modo que a análise de tal contexto fez parte do recorte da presente dissertação. Assim, a pesquisa se justifica pela relevância desse cenário em alterar a própria prática jurídica e afetar os direitos fundamentais dos jurisdicionados.

Como o assunto escolhido está em constante movimento, a escrita pode se tornar desafiadora pela necessidade de recorrentes atualizações, uma vez que a pesquisa em IA é uma atividade contínua e instigante. Assim, a investigação foi feita a partir do mapeamento das produções existentes, mas sem perder de vista que futuras publicações serão necessárias para acompanhar a evolução da temática.

O presente trabalho está dividido em quatro capítulos. Em razão da manifesta interdisciplinaridade do tema, o primeiro capítulo apresentou noções e conceitos sobre o que é a inteligência artificial, visto que há múltiplas definições para o termo e os estudiosos analisam o campo sob diferentes perspectivas. Percorreu-se um breve histórico da área, de maneira a retratar as oscilações de expectativas que marcaram as pesquisas nas últimas décadas. Tópicos relacionados aos algoritmos e ao aprendizado de máquina foram incluídos devido a sua importância para a compreensão do treinamento e da identificação de padrões a partir de dados pelos sistemas de IA. Na sequência, explicou-se as redes neurais artificiais e o subcampo do *deep learning*, responsáveis pelo avanço da IA em atividades mais complexas.

A IA generativa trouxe sofisticação aos modelos e alcançou popularidade com as ferramentas disponibilizadas ao público, capazes de gerar imagens, textos, áudios e vídeos. Assim, elucidou-se conteúdos atinentes a esse novo subcampo da IA, cujo potencial ainda está sendo explorado. A primeira parte da pesquisa, apesar de envolver nomenclatura mais voltada à ciência da computação, é essencial para uma melhor assimilação dos capítulos seguintes, que correlacionam esses pontos a outros conceitos.

O cerne do segundo capítulo são as relações da inteligência artificial e o Direito. Foram verificadas as transformações no mercado jurídico tendentes a incorporar a tecnologia em uma

lógica de apoio. Além disso, considerando a conjuntura que se iniciou em 2018 com o Projeto Victor, no Supremo Tribunal Federal, observa-se a crescente adoção de ferramentas de inteligência artificial no Poder Judiciário para otimizar os fluxos de trabalho e aprimorar a prestação jurisdicional. As análises realizadas têm como fonte indispensável as estatísticas produzidas pelo Conselho Nacional de Justiça, que evidenciam a utilização de tecnologia pelos tribunais.

Regular os sistemas de IA por meio de lei em sentido estrito é um debate que já dura alguns anos no Brasil. Os esforços do Poder Legislativo recaem sobre dezenas de Projetos de Lei apresentados tanto na Câmara dos Deputados quanto no Senado Federal, havendo a expectativa de aprovação de um marco legal que traga segurança e robustez para o desenvolvimento da área. Diante dessa lacuna, analisou-se atos normativos como a Estratégia Brasileira de Inteligência Artificial, instituída pela Portaria MCTI n. 4.617/21, e a Resolução CNJ n. 332/2020, sendo a última um fundamental direcionador ao Poder Judiciário.

O terceiro capítulo é dedicado aos fundamentos éticos da IA, visto que o seu potencial transformador em diferentes domínios provoca reflexões acerca dos princípios que orientam o seu uso e desenvolvimento. Os desafios e inquietações que permeiam o estudo foram examinados, com foco principalmente na questão do viés, que é indicada com profusão pela literatura. A implementação da ética pode ser problemática, por isso, na última seção do capítulo descreveu-se formas pelas quais é possível chegar-se à concretização desse ideal.

Através dos projetos de Pesquisa e Desenvolvimento (P&D) é possível consolidar e sistematizar conhecimentos e utilizá-los para a criação de aplicações, a exemplo das ferramentas de IA. O Projeto Sabiá, fruto de uma parceria entre a Universidade de Brasília e o Tribunal Superior Trabalho, está sendo desenvolvido para auxiliar o Tribunal a enfrentar os desafios na gestão do acervo e proporcionar uma melhor prestação jurisdicional mediante o uso de inteligência artificial. Por meio dos módulos iSimilares e iJulgados, em que se busca o agrupamento de processos por similaridade e o levantamento de jurisprudência, respectivamente, objetiva-se otimizar esforços e complementar o sistema interno Bem-Te-Vi, utilizado no gerenciamento de processos.

O último capítulo, portanto, é voltado a uma investigação empírica de um sistema de inteligência artificial destinado ao Poder Judiciário. Essa verificação é central para responder ao problema de pesquisa que orientou e delimitou o trabalho, consubstanciado na seguinte pergunta: “De que maneira o Projeto de Pesquisa e Desenvolvimento Sabiá está relacionado

aos parâmetros éticos da IA?". Embora não tenha sido concluído e esteja em fase de criação, o Projeto em questão pode ser analisado à luz da ética da IA. Para isso, foram utilizados os referenciais éticos desenvolvidos no capítulo anterior e as informações disponibilizadas no "Painel da Pesquisa sobre Inteligência Artificial 2023", divulgado pelo CNJ.

Quanto à metodologia, o procedimento utilizado é o dedutivo, em que se busca uma conclusão particular partindo-se de conhecimentos gerais; com o emprego da abordagem qualitativa, o que envolveu a análise exploratória de conteúdos sobre o tema. Aplicou-se as técnicas de pesquisa bibliográfica e documental.

1 O QUE É A INTELIGÊNCIA ARTIFICIAL

Russell e Norvig defendem que o ser humano, cuja espécie é denominada Homo sapiens – homem sábio – é caracterizado por sua inteligência. Este, durante milhares de anos, procura entender como pensa e age, isto é, compreender como um mero punhado de matéria pode perceber, prever e manipular um mundo muito maior e mais complicado que ele próprio. O campo da inteligência artificial, ou IA, vai além: tenta não apenas compreender, mas também construir entidades inteligentes que conseguem oferecer respostas a uma grande variedade de situações (Russell; Norvig, 2022, p. 14).

Inteligência Artificial é um termo da moda que tem sido internalizado pelo público em geral como uma noção abrangente, criando a impressão de que existe apenas uma entidade - a IA que pode ser abordada unitariamente. No entanto, considerando-se as tecnologias desenvolvidas e classificadas como aplicações de IA, essa percepção de unidade tende a desaparecer (Bertolini, 2020, p. 15-16).

Existem dispositivos tão diversos entre si, como um carro sem condutor, uma escova de dentes inteligente ou um sistema especializado para apoio em diagnóstico médico, que utilizam sistemas baseados ou operados por IA. Assim, uma compreensão desta noção é necessária para ajudar o público em geral a entender o que essas tecnologias implicam, o que podem provocar e como podem afetar o seu modo de vida e os seus direitos (Bertolini, 2020, p. 16).

Os pesquisadores têm definido a IA a partir de diferentes pontos de vista, alguns a definem em termos de fidelidade ao desempenho humano, enquanto outros preferem uma definição mais abstrata e formal, aproximando-se da chamada racionalidade (Russell; Norvig, 2022, p. 14). Fato é que inexiste consenso acerca de uma definição universal, coexistindo múltiplas acepções para o termo.

No âmbito do Laboratório de Pesquisa DR.IA[1], na Universidade de Brasília, foi desenvolvida uma noção ampliada de inteligência artificial, que leva em consideração a importância definidora dos dados nos seus alcances e impactos. No intitulado conceito “Lego” de IA, em referência ao brinquedo de blocos, a inteligência artificial é definida como “sistemas que buscam a reprodução parcial da atividade cognitiva realizada por seres humanos com o

[1] DR.IA - Laboratório de Direito e Inteligência Artificial. As publicações, notícias e atividades do Laboratório podem ser acessadas em: http://www.dria.unb.br/.

arranjo indispensável de três elementos: *dataset*, combinação algorítmica e resutados aferíveis" (Bonat; Hartmann Peixoto, 2023, p. 7-8).

Os elementos que compõem a definição "Lego" de IA devem ser devidamente combinados para que uma IA robusta seja compatível com as diretrizes de confiabilidade e respeito, de modo adequado a um tratamento jurídico protetivo aos direitos fundamentais. Essa visão ainda preconiza o envolvimento de ações de concretização de princípios éticos a cada evolução do desenvolvimento ou uso de IA (Bonat; Hartmann Peixoto, 2023, p. 8).

Para Jaime Simão Sichman (2021, p. 38), a IA é um ramo da ciência/engenharia da computação que visa desenvolver sistemas computacionais que solucionam problemas. Para tal, utiliza um número diverso de técnicas e modelos, dependendo dos problemas abordados. Diante disso, na visão do autor, é incorreto utilizar-se de expressões como "a IA da empresa X", sendo mais adequado, portanto, dizer " um sistema da empresa X que utiliza técnicas de IA".

Para Bertolini (2020, p. 9), uma grande parte da investigação em IA visa desenvolver soluções específicas, com funções bem definidas a serem operadas em determinadas situações (*light AI*), sendo que apenas uma pequena parte da investigação dedica-se à compreensão da IA como máquinas e softwares com capacidades e inteligência semelhantes às humanas (*general AI*). O autor ressalta que a única consideração fundamental e universal possível sobre os sistemas de IA é que não existe base filosófica, tecnológica ou jurídica para considerá-los outra coisa senão artefatos gerados pelo intelecto humano e, portanto, produtos.

A Estratégia Brasileira para a Transformação Digital (E-Digital) designa a IA como o conjunto de ferramentas estatísticas e algoritmos que geram softwares inteligentes especializados em determinada atividade. Aponta-se que se trata de tecnologia especialmente útil para classificação de dados, identificação de padrões e realização de predições. A estratégia assinala que ferramentas de tradução, serviços de reconhecimento de voz e imagens e mecanismos de buscas que ranqueiam sites de acordo com a relevância para o usuário são exemplos de amostras que se servem da IA (E-Digital. 2018, p. 61).

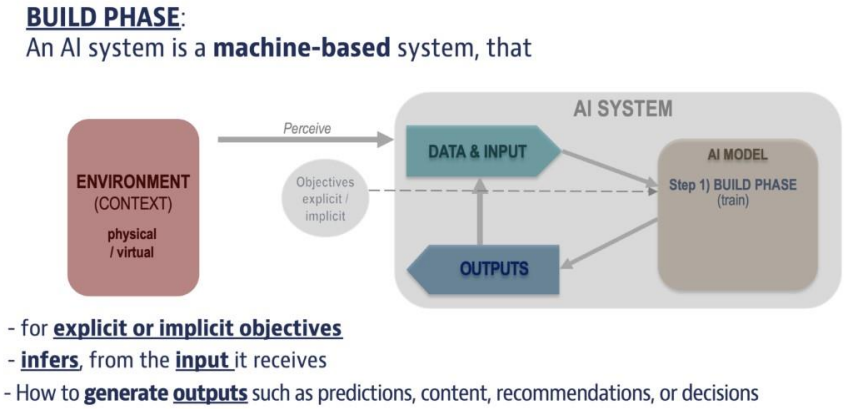
O fato de não existir uma definição consensual de Inteligência Artificial é reforçado pela Estratégia Brasileira de Inteligência Artificial (EBIA), instituída em 2021, que dispõe que esta é melhor entendia como um conjunto de técnicas destinadas a emular alguns aspectos da cognição de seres vivos usando máquinas (EBIA, 2021, p. 8).

Vale ressaltar que as definições de IA não ficam estagnadas no tempo. No âmbito da OCDE - Organização para a Cooperação e Desenvolvimento Econômico, por exemplo, foi lançada em 2023 uma atualização para o que seria um "sistema de IA". Para a Organização tais sistemas podem ser definidos como:

> Um sistema de IA é um sistema baseado em máquina que, para objetivos explícitos ou implícitos, infere, a partir das informações que recebe, como gerar resultados como previsões, conteúdos, recomendações ou decisões que podem influenciar ambientes físicos ou virtuais. Diferentes sistemas de IA variam nos seus níveis de autonomia e adaptabilidade após a implantação (Russell, Perset, Grobelnik, 2023)[2].

Em consonância com o conceito acima reportado, a fase de construção de um modelo, pode ser assim sintetizada:

Figura 1 - Modelo de sistema de IA da OCDE: fase de construção



Fonte: Russell, Perset, Grobelnik, 2023.

A atualização promovida pela OCDE procura refletir a ideia de que os objetivos de um sistema de IA podem ser explícitos ou implícitos. Enquanto os primeiros são caracterizados pela programação direta do objetivo por um desenvolvedor humano, nos implícitos o objetivo é atingido através de um conjunto de regras que são especificadas ou quando o próprio sistema é capaz de assimilar novos objetivos. Os sistemas autônomos de direção, por exemplo, são

[2] No original: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems in their levels of autonomy and adaptiveness after deployment".

treinados para cumprir as regras de trânsito, mas não “conhecem” seu objetivo implícito de proteger vidas (Russell, Perset, Grobelnik, 2023).

Houve a adição da palavra “conteúdo” no conceito da OCDE a fim de englobar os sistemas de IA generativos, que produzem “conteúdo”, como textos, vídeos ou imagens. Portanto, a revisão da definição é substantivamente importante para refletir os avanços nos desenvolvimentos tecnológicos (Russell, Perset, Grobelnik, 2023).

A cobertura midiática da IA distorce e muitas vezes exagera o seu potencial, tanto nos extremos positivos quanto nos negativos. Esse quadro fático favoreceu a sensibilização do público para as preocupações legítimas sobre vieses da IA, falta de transparência, prestação de contas e o potencial de automações guiadas por IA contribuir para a crescente desigualdade social (Littman et al., 2021, p. 8).

Uma boa compreensão sobre a IA possibilita que se desafie os relatos distópicos baseados na ficção científica que reduzem a confiança na tecnologia, atrasando a sua adoção mesmo quando desejável. Além disso, tal situação favorece um debate consciente sobre questões mais sensíveis, relativas a direitos fundamentais dos indivíduos (Bertolini. 2020, p. 16).

De acordo com Kahneman, Sibony e Sunstein (2022, p. 128-129), grandes conjuntos de dados são essenciais para análises sofisticadas, e a disponibilidade deles é uma das principais causas do rápido progresso da IA em anos recentes. Segundo os autores, o que a IA faz não envolve mágica nem compreensão, é mera identificação de padrões.

Diante de tantos avanços na área, escrever sobre o tema pode ser uma tarefa árdua e ingrata, apontam Księżak e Wojtczak (2023, p. 8), pois a noção de IA é um alvo em movimento, mudando ela própria com a evolução tecnológica, de forma que o que é considerado uma inovação hoje, em um momento futuro se tornará padrão, ou até obsoleto (Bertolini, 2020, p. 76).

Segundo Hartmann Peixoto e Silva (2019, p. 19), a IA tem como característica ser uma atividade multidisciplinar e, por essa razão, a intenção de delimitá-la implica em necessários recortes. O presente trabalho tem como delineamento a utilização da IA pelo Direito, com foco mais específico nos sistemas em uso pelo Poder Judiciário e nos princípios e diretrizes éticas que buscam assegurar o desenvolvimento e uso adequado de tais sistemas. Antes de iniciar esses tópicos, será feita uma contextualização sobre a área, com o estabelecimento de alguns conceitos-chave, o que permitirá uma melhor compreensão dos temas principais.

1.1 Breve histórico da inteligência artificial

Alan Turing (1912-1954), cujo trabalho essencial foi marcante para o futuro da IA, propôs em 1950 um teste com o objetivo de responder se uma máquina poderia pensar. Um computador passaria no experimento se o interrogador humano, depois de propor algumas perguntas por escrito, não conseguisse descobrir se as respostas às perguntas vieram de uma pessoa ou de um computador (Russell; Norvig, 2022, p. 15).

Para passar no teste, o computador precisaria ter capacidade de processamento de linguagem natural, de representar o conhecimento para armazenar o que sabe, de raciocínio automatizado para responder a perguntas e tirar novas conclusões e de aprendizado de máquina, para se adaptar a novas circunstâncias e detectar padrões. Para Turing, a simulação física de uma pessoa seria desnecessária para demonstrar inteligência. Muitos pesquisadores consideraram que estudar as capacidades da máquina que estão relacionadas ao teste e que são elementos essenciais do campo da IA, seria mais importante que o esforço à aprovação no teste (Russell; Norvig, 2022, p. 15).

Quase um século antes dos estudos Alan de Turing, a inglesa Ada Lovelace (1815-1852), mundialmente conhecida por seu pioneirismo nas ciências exatas, realizou trabalhos e escritos que versavam sobre a Máquina Analítica de Charles Babbage, projetada em 1837. Enquanto Babbage concebeu uma máquina programável de uso geral para o cálculo, Lovelace vislumbrou as infinitas possibilidades de estender a computação matemática a aplicações do mundo real (Reeve, 2019).

Alguns pesquisadores consideram Ada a primeira programadora da história, em razão da sua capacidade de desenvolver algoritmos que permitiram à máquina de Babbage computar os valores de funções matemáticas, e tal invenção possuía arquitetura precursora à dos computadores atuais (Martins, 2016, p. 14). O trabalho de Lovelace permaneceu em relativa obscuridade até a década de 1950, quando Alan Turing reconheceu sua contribuição visionária para o desenvolvimento da computação (Reeve, 2019).

De acordo com Octavia Reeve (2019), o legado de Lovelace aponta para a importância do pensamento interdisciplinar e demonstra que o trabalho intelectual colaborativo permite a construção de novos conhecimentos sobre bases coletivas, fortalecendo e diversificando as discussões sobre as tecnologias emergentes, o que inclui a inteligência artificial.

O termo inteligência artificial foi cunhado pelo matemático John MCarthy, em 1956, que convidou um grupo de estudiosos para participar de um *workshop* intitulado "*Dartmouth Summer Research Project on Artificial Intelligence*", no Dartmouth College, em Hanover, New Hampshire (Bertolini, 2020, p. 18). Apesar do lançamento da área sem um acordo acerca da metodologia, escolhas de problemas de pesquisa ou teoria geral, James Moor aponta que foi compartilhada a visão de que os computadores poderiam ser feitos para executar tarefas inteligentes (Moor, 2006, p. 87).

Segundo Haenlein e Kaplan (2019, p. 7), o evento no Dartmouth College, patrocinado pela Fundação Rockefeller, reuniu aqueles que mais tarde seriam considerados os pais fundadores da IA. Entre os participantes, estava o cientista da computação Nathaniel Rochester, projetista do IBM 701, o primeiro computador científico comercial e o matemático Claude Shannon, que fundou a teoria da informação. Esse período de nascimento da área resultou na chamada *AI Spring* (primavera da IA).

Durante o seminário de 1956, já havia a proposta de que a IA se tornasse um campo separado, e isso se deve ao fato de que esta abraçou desde o início a ideia de reproduzir faculdades humanas como a criatividade, o auto aperfeiçoamento e o uso da linguagem, sendo que até então nenhuma das outras áreas tratava dessas questões. Além disso, a IA é o único campo a tentar construir máquinas que funcionam de forma autônoma em ambientes complexos e mutáveis (Russell; Norvig, 2013, p. 39).

O seminário de Dartmouth foi seguido por um período de quase duas décadas de um sucesso significativo no campo da IA. Entre 1964 e 1966, foi criado o famoso programa de computador ELIZA por Joseph Weizenbaum no MIT. ELIZA foi uma ferramenta capaz de simular conversas com um humano e um dos primeiros programas a tentar passar no teste de Turing. No mesmo período, o vencedor do prêmio Nobel Herbert Simon, juntamente com cientistas da Rand Corporation, criou o programa General Problem Solver, que foi capaz de resolver automaticamente certos tipos de problemas simples, como as Torres de Hanói (Haenlein; Kaplan, 2019, p. 7).

Como resultado de histórias de sucesso inspiradoras, foi dado financiamento substancial à pesquisa em IA, e consequentemente, foram surgindo mais e mais projetos, culminando no chamado *AI summer* (verão da IA). Contudo, em 1973, o Congresso dos EUA começou a criticar fortemente os altos gastos com pesquisas na área; somado a isso, surgiram relatórios que questionavam a perspectiva otimista dada por pesquisadores sobre a nova tecnologia. Nesse

período, o governo britânico encerrou o apoio à pesquisa em inteligência artificial nas universidades, exemplo que foi seguido pelo governo dos EUA (Haenlein; Kaplan, 2019, pp. 7-8).

Na década de 1980, a teoria da probabilidade tornou-se a ferramenta dominante, em grande parte devido ao desenvolvimento das redes Bayesianas por Judea Pearl e outros cientistas da computação. Isso levou ao desenvolvimento das primeiras ferramentas computacionais em grande escala para raciocínio probabilístico, resultando em uma fertilização cruzada substancial entre IA e outros campos que se baseiam na teoria da probabilidade, incluindo estatística, teoria da informação, teoria de controle e pesquisa operacional (Russell, 2022, p. 45).

O estudo da inteligência artificial sempre foi cercado de enormes expectativas, e em inúmeras vezes essas não foram completamente atingidas. A oscilação de humor com relação ao campo é caracterizada por altos e baixos, havendo períodos de grande entusiasmo e grande financiamento (como ocorre agora) seguidos por outros de decepção e recursos escassos. Esses últimos são conhecidos como *AI winter* (Inverno da IA), como foram por exemplo os períodos entre 1975/1980 e 1987/1993 (Sichman, 2021, p. 37).

Para além do que foi exposto, Russell e Norvig (2022, p. 15) apontam que uma forma rápida de resumir os marcos na história da IA é listar os vencedores do Prêmio Turing: Marvin Minsky (1969) e John McCarthy (1971) pela definição dos fundamentos do campo com base na representação e no raciocínio; Ed Feigenbaum e Raj Reddy (1994) pelo desenvolvimento de sistemas especialistas, que codificam o conhecimento humano para resolver problemas do mundo real; Judea Pearl (2011) pelo desenvolvimento de técnicas de raciocínio probabilístico; e Yoshua Bengio, Geoffrey Hinton e Yann LeCun (2018) por tornar o “aprendizado profundo” uma parte essencial da computação moderna.

1.2 Algoritmos e a inteligência artificial

Um algoritmo nada mais é do que uma sequência finita de ações que resolve um certo problema. Uma receita culinária, como a de um risoto, é um algoritmo. Desse modo, um algoritmo pode resolver problemas de tipos bastante diferentes: cálculos para o projeto de uma ponte, processamento de dados para a geração de uma folha de pagamento ou até mesmo o planejamento para a definição de um pacote de turismo (Sichman, 2021, p. 38).

Boa parte das rotinas das pessoas baseia-se no contato com algoritmos, pois estes determinam os *feeds* de notícias, as buscas na internet, as sugestões e influências do marketing digital e, de maneira não tão perceptível, estão presentes na previsão do tempo, na seleção de currículos para uma vaga de emprego, na organização de pautas de discussão social e nas influências políticas (Hartmann Peixoto; Silva, 2019, p. 69).

Kahneman, Sibony e Sunstein (2022, p. 123) indicam que um algoritmo pode ser definido como um processo ou série de regras a serem seguidas em cálculos ou outras operações de resoluções de problemas, especialmente por um computador. Hartmann Peixoto e Silva (2019, p. 73) afirmam que um algoritmo pode ser descrito de várias formas, o que inclui a linguagem natural e até mesmo representações gráficas em um fluxograma, de modo a buscar sempre o caminho mais objetivo para a finalidade pretendida.

É preciso elucidar que não se deve compreender os algoritmos como sinônimo de inteligência artificial. As tarefas desempenhadas por sistemas de IA diferenciam-se de forma significativa das tarefas tradicionais de algoritmos (tais como classificar listas de números ou calcular raiz quadrada). Sistemas que utilizam inteligência artificial, por sua vez, são compostos de algoritmos juntamente com outras aplicações de computadores (Hartmann Peixoto; Silva, 2019, p. 76; p. 84).

Após a criação de um algoritmo para um sistema de IA, explica Lage, a próxima etapa é o seu treinamento, que consiste no estágio no qual são fornecidos casos de amostra e de resultado esperado. O conjunto de informações concedidas deve ser grande o suficiente e estatisticamente significativas para fazer sentido para o sistema fornecer um resultado razoável, preciso e previsível. Por fim, aquele que interpreta o resultado deve ser capaz de seguir o caminho que levou à conclusão e obter informações sobre a precisão (Lage, 2021, p. 55).

Os algoritmos, aliados às máquinas cada vez mais potentes e sofisticadas, representam um motor do desenvolvimento científico em todos os campos do conhecimento humano. Contudo, existe a preocupação com possíveis efeitos danosos e desregrados que a utilização dos algoritmos pode causar (Lage, 2021, p. 55).

O grande desafio para a criação de um sistema de IA é a consciência acerca de sua definição, ou seja, utilizar algoritmos apropriados para solucionar um problema ou conjunto de problemas específicos. Não há design universal para criar um modelo, mas organizações de conexões específicas para problemas específicos (Hartmann Peixoto; Silva, 2019, p. 64).

1.3 Aprendizado de máquina e os objetivos da IA

A inteligência artificial, através dos métodos de aprendizado de máquina, possibilitou aos computadores realizar tarefas que eram tidas como essencialmente humanas, como reconhecer rostos e analisar imagens radiológicas. Kahneman, Sibony e Sunstein (2022, p. 123) apontam que algoritmos de aprendizado de máquina são capazes executar previsões complexas, como prognósticos de decisões da Suprema Corte Americana e avaliação de quais denúncias ao serviço de proteção infantil americano exigem a verificação mais urgente de um assistente social.

O aprendizado de máquina, ou *machine learning*, está associado ao processamento de um considerável volume de dados para a identificação de padrões que, também combinados, possibilitam a predição e recomendação de ações características da atividade cognitiva humana (Hartmann Peixoto, 2020a, p. 18). Portanto, algoritmos de aprendizado de máquina são capazes de encontrar padrões a partir combinações de variáveis que poderiam passar despercebidas ou serem negligenciadas pelo olhar humano (Kahneman; Sibony; Sunstein, 2022, p. 130).

Segundo Vishal Maini e Samer Sabri (2017, p. 9), o aprendizado de máquina é um subcampo da inteligência artificial cujo objetivo é permitir que os computadores aprendam por conta própria. Um algoritmo de *machine learning* possibilita a identificação de padrões em dados observados, a construção de modelos que expliquem o mundo e prevejam coisas sem a necessidade de regas e modelos explícitos pré-programados.

A divisão do aprendizado de máquina em três tipos é feita por alguns autores, que o classificam em supervisionado, não supervisionado e por reforço (Hartmann Peixoto; Silva, 2019, p. 91). O aprendizado supervisionado é utilizado nas tarefas de regressão (predição de um valor numérico contínuo) e de classificação (atribuição de um rótulo). A predição da renda de uma pessoa com base no número de anos de ensino superior é um exemplo que envolve a regressão, enquanto que um algoritmo que identifique se uma imagem ilustra um gato ou um cachorro está situado na tarefa de classificação (Maini; Sabri, 2017, p. 16-19).

O funcionamento de um algoritmo de aprendizado supervisionado inicia-se a partir de um conjunto de dados (*dataset*) que contenha exemplos de treinamento associados a rótulos (*labels*) corretos. Executando os dados de treinamento rotulados, a máquina tenta aprender a relação pretendida, para em sequência aplicar o modelo aos dados de teste não rotulados. O objetivo da aprendizagem supervisionada é exercer a regressão ou classificação desejada com

a maior precisão possível com novos exemplos como entrada - *input* (Maini; Sabri, 2017, p. 17-18).

Um filtro de *spam*, através do aprendizado de máquina, pode aprender a marcar os e-mails recebidos como *spam* ou como não-*spam*. Utilizando um conjunto de treinamento, em que cada exemplo se chama instância de treinamento (ou amostra), é possível treinar o programa para sinalizar se os novos e-mails são mensagens indesejadas. A medida de desempenho, que pode ser definida a partir da proporção de e-mails sinalizados corretamente, chama-se acurácia e é frequentemente usada em tarefas de classificação (Aurélien, 2021, p. 16). Também é factível avaliar o desempenho por meio do cálculo da proporção de exemplos para os quais o modelo produz uma saída incorreta, a fim de se chegar à taxa de erro (Cozman; Kaufman, 2022, p. 197).

Hartmann Peixoto e Silva (2019, p. 93) alertam que a precisão da classificação ou da regressão dependerá fatores como a efetividade do algoritmo escolhido, de como ocorre a sua aplicação e da quantidade e qualidade dos dados usados para treinamento.

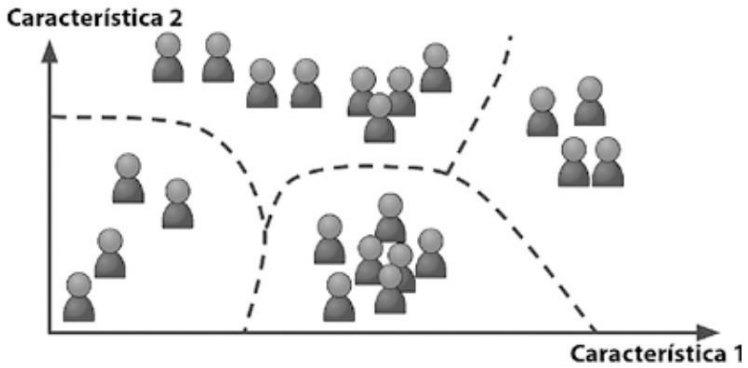
Algoritmos de aprendizagem não supervisionada, por sua vez, não recebem dados rotulados. O Agrupamento dos dados por similaridade (*clustering*) e a redução da dimensionalidade para compactar os dados, mantendo sua estrutura e utilidade (*dimensionality reduction*) são tarefas de destaque nesse tipo de aprendizagem. Nem sempre é fácil criar métricas sobre a performance de um algoritmo de aprendizado não supervisionado, pois o desempenho é frequentemente subjetivo e específico de um domínio (Maini; Sabri, 2017, p. 55-56).

A redução de dimensionalidade é frequentemente utilizada para pré-processar os dados, compactando-os de uma maneira que preserve o seu significado, antes de alimentá-lo em uma rede neural profunda ou em outro algoritmo de aprendizado supervisionado (Maini; Sabri, 2017, p. 57; p. 67).

O objetivo da clusterização é criar grupos de dados representados por pontos, de modo que os pontos distribuídos em diferentes grupos (*clusters*) não são similares, mas os pontos dentro de um mesmo *cluster* possuem similaridade (Maini; Sabri, 2017, p. 57). Um exemplo dessa classificação, fornecida por Aurélien Géron (2021, p. 19), consiste em mapear os visitantes de um blog a partir de dados disponíveis sobre eles. O próprio algoritmo encontrará conexões entre os visitantes, de modo que as relações observadas podem identificar que certa porcentagem dos visitantes são homens que adoram histórias em quadrinhos, enquanto uma

porcentagem menor representa mulheres amantes de ficção científica. Dessa forma, permite-se que as postagens sejam direcionadas para cada tipo de grupo.

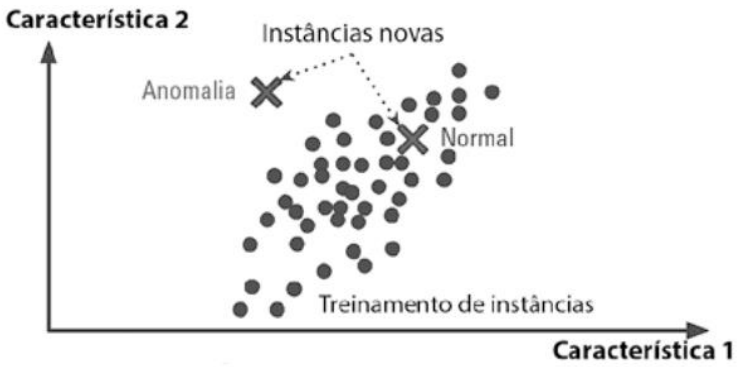
Figura 2 - Exemplo de clusterização



Fonte: Aurélien, 2021, p. 19.

Detectar anomalias também é uma tarefa fundamental exercida por algoritmos de aprendizagem não supervisionada. Essa atividade resulta na detecção de transações incomuns em cartões de crédito a fim de evitar fraudes, na identificação de defeitos de fabricação ou na remoção automática de *outliers* (dados que se diferenciam drasticamente dos restantes) de um conjunto antes de fornecê-lo a outro algoritmo de aprendizado (Aurélien, 2021, p. 19). A figura abaixo reproduzida ilustra a detecção de anomalia.

Figura 3 - Exemplo de detecção de anomalia



Fonte: Aurélien, 2021, p. 19.

Se os dados de treinamento estiverem cheios de erros, *outliers* e ruídos, o sistema terá mais dificuldade para detectar padrões básicos, o que diminui a chance do bom funcionamento do modelo. Além disso, o sistema só será capaz de aprender se os dados de treinamento tiverem características relevantes suficientes e poucas características irrelevantes. A seleção de um bom

conjunto de atributos para o treinamento, processo chamado de *feature engeneering* é imprescindível para o sucesso de um projeto de aprendizado de máquina (Aurélien, 2021, p. 23).

Aprender a partir da interação é uma ideia fundamental subjacente a quase todas as teorias de aprendizagem e inteligência, ressaltam Sutton e Barto (2015, p. 1). Para os autores, essa é a premissa da chamada aprendizagem por reforço, que é muito mais focada na aprendizagem direcionada a objetivos a partir da interação do que outras abordagens de aprendizagem de máquina.

Os problemas de aprendizagem por reforço envolvem aprender o que fazer de modo a maximizar um sinal numérico de recompensa. O sistema não é informado sobre quais ações tomar, como acontece em muitas formas de aprendizado de máquina, mas, em vez disso, deve descobrir quais ações geram mais recompensas ao experimentá-las. Nessa abordagem enfatiza-se a aprendizagem de um agente a partir da interação direta com seu ambiente, sem depender de supervisão exemplar ou de modelos completos do ambiente (Sutton; Barto, 2015, p. 2; p. 15).

Para uma melhor compreensão da aprendizagem por reforço, Sutton e Barto (2015, p. 5) fornecem o exemplo de um robô móvel que tem como função coletar lixo; em dado momento ele precisa decidir se deve entrar em uma nova sala em busca de mais lixo para coletar ou se deve encontrar o caminho de volta para a estação de recarga de bateria. Ele toma sua decisão com base no nível de carga atual de sua bateria e na rapidez e facilidade com que conseguiu encontrar o carregador no passado. Assim, esse agente procura atingir um objetivo apesar da incerteza sobre o seu ambiente e utiliza a sua experiência para melhorar o seu desempenho ao longo do tempo (Sutton; Barto, 2015, p. 8).

Outro critério utilizado para classificar os sistemas de aprendizado de máquina é se estes podem ou não aprender de forma incremental a partir da entrada de um fluxo de dados. No aprendizado em *batch* (por ciclo), o sistema é incapaz de aprender de forma incremental. Assim, para que o sistema tenha acesso a dados novos, é preciso treinar uma nova versão a partir do zero, inserindo os dados novos e antigos, e, em seguida, descontinuar o sistema antigo e substituí-lo pelo novo. Esse tipo de treinamento demanda muito tempo e exige mais recursos computacionais (Aurélien, 2021, p. 20).

Com um custo mais baixo e de forma mais rápida, no aprendizado online (incremental) é possível treinar o sistema fornecendo as instâncias de dados de forma sequencial, individual

ou em pequenos grupos, chamados de *mini-batches*. Nessa configuração, o sistema pode utilizar os dados novo em tempo real, tão logo eles entram. Esse tipo de aprendizado é excelente para sistemas que recebem dados em fluxo contínuo (por exemplo, preços das ações) e precisam se adaptar a mudanças de forma rápida ou autonomamente (Aurélien, 2021, p. 20).

Aurélien Géron (2021, p. 17) sintetiza que a aplicação do aprendizado de máquina é uma boa opção para lidar com problemas para os quais as soluções atuais exigem muitos ajustes finos ou extensas listas de regras e para problemas complexos nos quais não existe uma boa solução utilizando-se da abordagem tradicional. O autor reitera que além de ser capaz de operar com grandes quantidades de dados, um sistema de aprendizado de máquina pode ter a capacidade de se adaptar a novos dados.

Em razão de ser um sistema probalístico, com base na alimentação de dados, os algoritmos de *machine learning* podem, além dos benefícios, trazer problemas que refletem discriminações, parcialidades, escolhas ofensivas e desinformações. Portanto, é possível que o sistema computacional reflita os valores implícitos dos seus desenvolvedores, caracterizando o chamado viés humano - *human bias* (Hartmann Peixoto; Silva, 2019, p. 34-35).

1.4 As redes neurais artificiais e a sua relação com o *deep learning*

A lógica da estrutura cerebral inspirou a construção das redes neurais artificiais (RNAs), as quais estão no centro nevrálgico do aprendizado de máquina. As RNAs são versáteis, poderosas e escalonáveis, o que as tornam ideais para tarefas grandes e extremamente complexas de aprendizado de máquina, como classificar bilhões de imagens, habilitar serviços de reconhecimento de fala, recomendar vídeos para usuários ou aprender a vencer o campeão mundial no jogo Go - feito realizado pelo AlphaGo, da DeepMind (Aurélien, 2021, p. 117).

Segundo Aurélien Géron (2021, p. 117), uma arquitetura precursora das redes neurais artificiais foi apresentada pela primeira vez em 1943, pelo neurofisiologista Warren McCulloch e pelo matemático Walter Pitts. No artigo de referência destes estudiosos, exibiu-se um modelo computacional simplificado de como os neurônios biológicos podem trabalhar juntos na realização de cálculos complexos usando a lógica proposicional. A partir de então, outros modelos foram desenvolvidos.

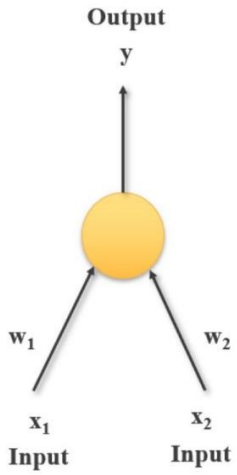
Na década de 60, os financiamentos na área foram cortados e as RNAs entraram em um longo inverno. No início dos anos 1980, foram inventadas novas arquiteturas e desenvolvidas melhores técnica de treinamento, o que suscitou um renascimento do interesse no conexionismo

(o estudo de redes neurais). No entanto, em meados dos anos 1990, outras técnicas foram inventadas e o estudo das redes neurais novamente foi deixado de lado. Atualmente as RNAs entraram em um círculo virtuoso de financiamento e progresso, em que produtos baseados nessas redes figuram em notícias que atraem cada vez mais atenção (Aurélien, 2021, p. 118).

Como uma rede neural é um modelo projetado para simular vagamente o processo de aprendizagem do cérebro humano, ela é concebida de forma que possa identificar padrões subjacentes nos dados e aprender com eles. Ressalta-se que é preciso converter quaisquer dados em um formato numérico antes de alimentá-los na rede neural, quer se trate de dados visuais, textuais ou séries temporais. Logo, é preciso representar problemas de forma que possa ser compreendido por essas redes (Artasanchez; Joshi, 2020, p. 469-470).

O perceptron é a estrutura mais básica de uma rede neural. Ele recebe entradas, realiza cálculos sobre elas e então produz uma saída, utilizando uma função linear simples para dar a resposta. Através do modelo foi possível obter uma melhor compreensão das RNAs. Contudo, para resolver problemas complexos, um modelo tão simples não é suficiente (Artasanchez; Joshi, 2020, p. 472 -475). Na figura abaixo é reproduzida a funcionalidade básica do Perceptron:

Figura 4 - Arquitetura básica de um perceptron

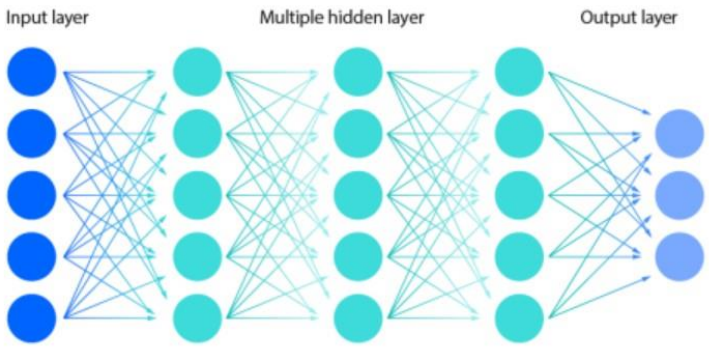


Fonte: Artasanchez; Joshi, 2020, p. 472.

Redes neurais são construídas usando camadas de unidades (ou nós), que são inspiradas nos neurônios biológicos. Cada camada de uma rede neural é um conjunto de unidades independentes conectadas às unidades da camada adjacente (Artasanchez; Joshi, 2020, p. 470). Uma rede neural possui uma camada de entrada, uma ou mais camadas ocultas e uma camada

de saída. Cada nó tem seu próprio peso e limite associados; se a saída de qualquer nó individual estiver acima do valor limite especificado, essa unidade será ativada, enviando dados para a próxima camada da rede. Caso contrário, nenhum dado será transmitido para a próxima camada da rede (IBM, [s.d]). A imagem a seguir ilustra uma arquitetura de RNA com múltiplas camadas:

Figura 5 - Rede Neural com múltiplas camadas ocultas



Fonte: IBM, [s.d].

As redes neurais podem ser classificadas em diferentes tipos, a depender do fim para o qual vai ser utilizada. Redes neurais feedforward, por exemplo, são compostas por unidades sigmoides, em que os dados inseridos são a base para a visão computacional e o processamento de linguagem natural. Nas redes neurais convolucionais (CNNs), princípios da álgebra linear são aproveitados para o reconhecimento de imagens e de padrões. As redes neurais recorrentes (RNNs) são identificadas por seus ciclos de feedback, no qual dados de séries temporais são usados para fazer previsões sobre resultados futuros (IBM, [s.d]).

O Aprendizado profundo (*deep learning*) é uma forma específica de aprendizado de máquina que envolve o treinamento de redes neurais com muitas camadas de unidades (Hartmann Peixoto e Silva, 2019, p. 99). Técnicas de *deep learning* são responsáveis pelo avanço substancial do estado da arte em reconhecimento visual de objetos, reconhecimento de fala e tradução automática, os quais são na opinião de Russell (2022, p. 46) os três dos subcampos mais importantes da tecnologia.

A ascensão de GPUs (unidades de processamento gráfico) e outros hardwares orientados para o *deep learning*, juntamente com o desenvolvimento de softwares de código aberto que facilitaram a expressão de modelos e cálculos, contribuíram para o impulso da aprendizagem profunda. Essas estruturas resultaram na aplicação desses avanços a uma variedade

incrivelmente ampla de domínios de problemas e na proliferação de pesquisas sobre o tema (Dean, 2022, p. 61-62).

Redes neurais de aprendizado profundo (*deep learning neural networks* - DLNN) possuem certas limitações, pois a arquitetura complexa dos modelos demanda grande capacidade de processamento computacional. Além disso, a qualidade dos resultados depende dos dados utilizados no desenvolvimento, treinamento e aperfeiçoamento dos sistemas. Assim, vieses podem emergir em função das decisões tomadas pelos desenvolvedores, em razão de erros na rotulagem da base de dados e também na própria geração de dados, a exemplo da não desagregação por gênero (Cozman; Kaufman, 2022, p. 198).

Outra questão que pode gerar inquietações é a utilização do aprendizado profundo para produzir inverdades que não são facilmente detectadas. Isso vem ocorrendo através do *deepfake*, palavra surgiu da combinação dos termos "*deep learning*" e "*fake*", e indica a utilização de softwares baseados em técnicas de aprendizado profundo para a produção de mídias falsas, a exemplo da substituição do rosto de uma pessoa por outra, em uma imagem ou vídeo, de modo a imitar as expressões faciais, a voz e as inflexões do indivíduo substituído. Os conteúdos *deepfake* representam uma grande ameaça à privacidade, à segurança social e à integridade na Internet (Chadha et al., 2021, p. 558-564).

Segundo Chadha et al. (2021, p. 564), a melhoria significativa na qualidade e capacidade das redes generativas profundas torna o conteúdo *deepfake* mais realista e perfeito. No próximo tópico serão detalhadas questões acerca da IA generativa, que tem sido impulsionada pelos avanços na compreensão das redes neurais e do *deep learning* e vem sendo o centro das expectativas sociais.

1.5 A ascensão da IA generativa

Hartmann Peixoto e Silva (2019, p. 43) destacam que a IA não é um fenômeno novo, mas as tecnologias disponíveis permitem, diante da imensidão de dados produzidos a cada dia, visualizar uma nova dimensão desse feito. Recentemente, vem ganhando cada vez mais popularidade um novo tipo de modelo: é a chamada IA generativa, que se refere a sistemas que geram diversos tipos de dados, como imagens, textos, áudios e vídeos (Stripling, Abel, 2023).

Os modelos de IA discriminativa têm como objetivo prever a qual classe pertence uma instância, a exemplo de identificar se uma transação é fraudulenta ou legítima. Em contraste,

os modelos generativos geram a instância a partir do rótulo. Por exemplo, dado o comando “um gato tocando banjo”, é gerada a imagem de um gato tocando banjo. (Stripling; Abel, 2023).

Figura 6 - Imagem gerada pela ferramenta Craiyon, a partir do prompt “a cat playing banjo”

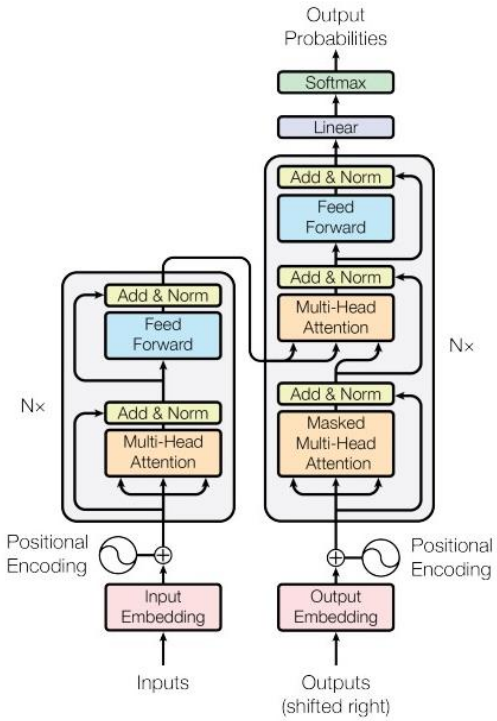
Fonte: Stripling; Abel, 2023.

Outra ferramenta semelhante ao Craiyon, é o DALL-E, da OpenAI, que também permite aos usuários descrever uma imagem usando texto, de modo que a figura a ser constituída possua as propriedades expressas pela descrição em linguagem natural. Esse tipo de *software*, através do uso de algoritmos de aprendizagem profunda, surpreende pelo nível de sofisticação dos resultados gerados e possibilita que artistas e outros criadores externalizem rapidamente suas ideias (Dean, 2022, p. 65).

O poder da IA generativa vem do uso dos chamados “*transformers*”, que produziram uma revolução em 2018 no processamento de linguagem natural (Stripling, 2023). Um artigo do Google Brain intitulado “*Attention is All You Need*”, publicado em 2017, é famoso por popularizar os conceitos de mecanismo de atenção e de “*transformer*”. Os autores do estudo mostram como é possível criar redes neurais poderosas, por meio da arquitetura dos “*transformers*”, para modelagem sequencial sem a necessidade de arquiteturas recorrentes ou convolucionais complexas, mas, em vez disso, utilizando mecanismos de atenção (Foster, 2023, p. 384-385).

O esquema abaixo representa o funcionamento de uma arquitetura “*tranformer”*. No lado esquerdo, um conjunto de blocos codificadores *transformer* codifica a sequência a ser traduzida. No lado direito, um conjunto de bloco decodificadores *transformer* gera o texto traduzido (Foster, 2023, p. 417).

Figura 7 - Modelo de arquitetura *transformer*



Fonte: Vaswani et al., 2017.

Um mecanismo de atenção é o que torna esse tipo de arquitetura única e distinta das abordagens recorrentes de modelagem de linguagem, pois este é capaz extrair informações úteis de maneira eficiente, sem ser obscurecido por detalhes irrelevantes. Isso o torna altamente adaptável a uma série de circunstâncias no momento da inferência (Foster, 2023, p. 388-389).

Quase um ano após a publicação do primeiro trabalho sobre o "*Transformer*", a OpenAI introduziu o modelo GPT (*generative pre-trained transformer*), em junho de 2018, no artigo intitulado "*Improving Language Understanding by Generative Pre-training*". No trabalho, os autores mostraram como uma arquitetura "*transformer*" pode ser treinada com uma enorme quantidade de dados de texto para prever a próxima palavra em uma sequência. Após o pré-treinamento, o modelo GPT pode ser ajustado para uma tarefa específica, como responder a perguntas. (Foster, 2023, p. 385-386).

Desde então, a arquitetura GPT foi aprimorada e ampliada pela OpenAI, que lançou modelos subsequentes, como as versões GPT-2, GPT-3, GPT-3.5 e GPT-4. Esses modelos foram treinados em conjuntos de dados maiores, para que gerassem textos mais complexos e coerentes. Os modelos GPT foram amplamente adotados por pesquisadores e profissionais,

contribuindo para avanços significativos nas tarefas de processamento de linguagem natural (Foster, 2023, p. 386).

A IA generativa tornou-se um importante tópico de discussão devido ao seu uso em chatbots, como o ChatGPT (da OpenAi) ou o Gemini (do Google). Por trás de produtos como estes estão os *Large Language Models*, ou LLMs. Esse termo é usado para modelos de linguagem treinados com grandes *datasets* que contêm quantidades massivas de parâmetros, a exemplo do GPT-3, subjacente a uma das versões do ChatGPT, que têm mais de 175 bilhões de parâmetros e um corpus de mais de meio trilhão de *tokens* (Stripling, Abel, 2023).

David Foster (2023, p. 628-629) elaborou tabela que resume alguns dos LLMs mais poderosos hoje existentes, com a indicação do modelo, data de lançamento, desenvolvedor, número de parâmetros e se tem o código aberto ou não.

Tabela 1 - Exemplos de Large Language Models

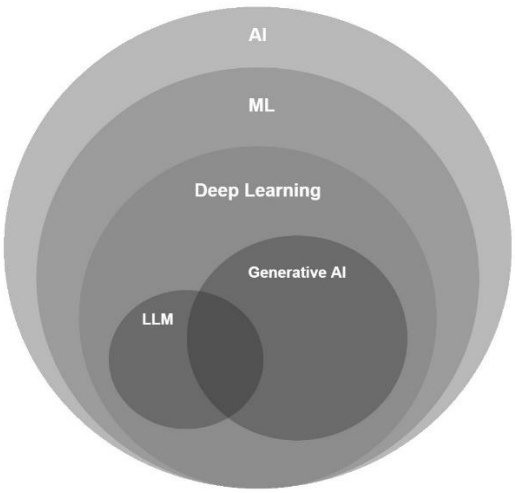
Model	**Date**	**Developer**	**#parameters**	**Open source**
GPT-3	May 2020	Open AI	175 billions	No
GPT-Neo	Mar 2021	Eleuther AI	2,7 billions	Yes
GPT-J	Jun 2021	Eleuther AI	6 billions	Yes
Megatron-Turing NLG	Oct 2021	Microsoft & NVIDIA	530 billions	No
Gopher	Dec 2021	DeepMind	280 billions	No
LaMDA	Jan 2022	Google	137 billions	No
GPT-NeoX	Fev 2022	Eleuther AI	20 billions	Yes
Chinchilla	Mar 2022	DeepMind	70 billions	No
PaLM	Apr 2022	Google	540 billions	No
Luminous	Apr 2022	Aleph Alpha	70 billions	No
OPT	May 2022	Meta	175 billions	Yes
BLOOM	Jul 2022	Hugging Face collaboration	175 billions	Yes
Flan-T5	Oct 2022	Google	11 billions	Yes
GPT-3.5	Nov 2022	OpenAI	Unknown	No

LLaMA	Feb 2023	Meta	65 billions	No
GPT-4	Mar 2023	OpenAI	Unknown	No

Fonte: Foster, 2023 (com adaptações).

Assim como a IA generativa, os grandes modelos de linguagem (LLM) são subconjuntos do *deep learning* e ambos possuem áreas de intersecção, o que significa a possibilidade de utilização desse tipo modelo para sistemas de IA que geram conteúdo. Um *Large Languade Model* é treinado para resolver problemas comuns de linguagem, como classificação de texto, respostas a perguntas, resumo de documentos e geração de texto, de forma que os modelos podem ser adaptados para resolver problemas específicos em diferentes áreas (Ewald, 2023).

Figura 8 - Intersecção entre IA generativa e os Large Languages Models (LLMs)



Fonte: Ewald, 2023.

Segundo Russell, Perset e Grobelnik (2023), um *Large Language Moldel* (LLM) é um exemplo de sistema com objetivo implícito, já que o objetivo não é explicitamente programado, mas adquirido em parte através do processo de aprendizagem por imitação de textos gerados por humanos e em parte através do processo de aprendizagem por reforço a partir de feedback humano - RLHF.

A aprendizagem por reforço a partir das preferências humanas (*reinforcement learning from human feedback* -RLHF) é especialmente adequada para tarefas com objetivos complexos, mal definidos ou difíceis de especificar. Em termos práticos, seria inviável para uma solução algorítmica definir algo "engraçado" de forma matemática, mas os humanos são capazes de avaliarem piadas consideradas engraçadas, por exemplo. Esse feedback humano, convertido em uma função de recompensa, pode ser usado para melhorar as habilidades de escrita de piadas de um grande modelo de linguagem (LLM) (Bergmann, 2023).

Existem limitações advindas do uso de RLHF tais como: os dados de preferência humanas são caros; o input humano é altamente subjetivo; os avaliadores humanos podem falhar, ou até mesmo ser intencionalmente adversários e maliciosos; há o risco de propagação de vieses, caso o feedback humano seja coletado a partir de um grupo demográfico muito restrito e a sua utilização se dê no âmbito de diferentes grupos (Bergmann, 2023).

Modelos de linguagem baseados em redes neurais muito profundas, que são treinados para prever a próxima palavra em uma sequência, mostram habilidades intrigantes para responder a perguntas de uma forma semanticamente significativa. Contudo, pesquisas evidenciam que os sistemas de *deep learning* muitas vezes não conseguem generalizar de forma robusta e são suscetíveis a indicar regularidades espúrias com base nos dados de treinamento (Russell, 2022, p. 46).

As chamadas alucinações (*hallucinations*) são limitações dos modelos que resultam em palavras ou frases absurdas ou gramaticalmente incorretas (Stripling, 2023). O problema é causado principalmente por dados de treinamento enviesados, comandos ambíguos e parâmetros de LLM imprecisos, e ocorrem principalmente ao combinar fatores matemáticos com contexto baseado em linguagem. Assim, ao projetar soluções baseadas em linguagem, é necessário monitorar e controlar as alucinações a fim de evitar repercussões indesejadas (Roychowdhury, 2023).

Conforme visto, os grandes modelos de linguagem têm o propósito de modelar diretamente a linguagem a partir de um enorme corpus de texto (Foster, 2023, p. 625). Entretanto, a fonte dos dados de treinamento tem provocado crescentes disputas. O New York Times, por exemplo, processou a OpenAI e a Microsoft, no final de 2023, pela utilização indevida de milhões de artigos para treinamento de *chatbots* baseados em IA generativa. Entre outros pontos, a ação afirma que o *chatbot* forneceu aos usuários trechos quase literais de artigos

do Times que, de outra forma, exigiriam uma assinatura paga para visualização (Grynbaum; Mac, 2023).

Stripling e Abel (2023) alertam que quanto mais complexos os modelos se tornam, mais difícil são de serem compreendidos. De acordo com os autores, mesmo para uma rede neural com apenas uma camada oculta, é difícil descrever a conexão entre os elementos utilizados. Isso se torna ainda mais difícil com os modelos muito grandes usados para problemas com dados não estruturados e modelos generativos.

2 INTELIGÊNCIA ARTIFICIAL E O DIREITO

O mercado jurídico brasileiro é vasto e há um déficit informacional acerca da atual realidade das ocupações jurídicas, apontam Bonat e Hartmann Peixoto (2023, p. 19-20). De acordo com os autores, houve, nas últimas décadas, uma transformação intensa no campo da atuação jurídica que não é refletida na denominada Classificação Brasileira de Ocupações - CBO, desenvolvida a partir da década de 80 pelo Ministério do Trabalho e Emprego - MTE.

Em pesquisa divulgada em 2021, pelo Instituto de Pesquisa Econômica Aplicada- IPEA, sobre o mercado de trabalho jurídico no Brasil, foi relatado que em 2020 havia quase um milhão de ocupados como profissionais jurídicos. Destacou-se que o mercado é voltado essencialmente à prestação de serviços para o setor privado da economia, com uma variabilidade acentuada da remuneração entre os trabalhadores (Campos; Benedetto, 2021, p. 17-19).

Diante de uma área tão abrangente como é o caso do Direito, refletir sobre o uso de soluções baseadas em IA envolve analisar aspectos como as ocupações que fazem parte da rotina dos profissionais, as habilidades requeridas no trabalho desempenhado e até mesmo investigar o quão sensível são as atividades jurídicas ao uso dados (Bonat; Hartmann Peixoto, 2023, p. 19-20).

A IA contribui para um movimento disruptivo no tradicional mercado do Direito. Há alterações de estratégias de escritórios de advocacia, tanto no aspecto de estruturação interna como na atuação no contencioso e no consultivo. Outro movimento são as influências da IA nos entes governamentais de maneira geral e, especialmente no Poder Judiciário pelo impacto no tempo de duração de um processo. Há também a tendência de transformação da educação jurídica, mas que dependerá de um grau maior de maturidade sobre o tema (Hartmann Peixoto; Silva, 2019, p. 58-59).

Diversas *startups* no contexto da advocacia, denominadas Legaltechs ou Lawtechs, criam produtos e serviços de base tecnológica para melhorar o setor jurídico. Entre as ferramentas oferecidas estão: *chatbots*, inteligência artificial preditiva, *smart contracts*, sistemas de gerenciamento de caso e ferramentas de pesquisa inteligente. O desenvolvimento desse mercado foi possível em razão do avanço na modernização do Direito e da formação de novos advogados, que lidam com a tecnologia de forma mais natural que as gerações anteriores. (Lage, 2021, p. 113-115).

Segundo Susskind, os sistemas de IA são capazes de oferecer uma variedade de previsões notáveis relacionadas ao mundo jurídico. Na revisão de documentos extensos, a assistência pela tecnologia é capaz de superar a performance de advogados júniores e paralegais; em contratos, a partir de grandes bases de dados, os sistemas podem ser aptos a indicar aqueles que parecem dar origem aos riscos e responsabilidades mais críticas. Para o autor, as máquinas estão se tornando cada vez mais capazes e assumindo mais e mais tarefas que antes eram competência exclusiva dos seres humanos (Susskind, 2019, p. 271; p. 273).

A interdisciplinaridade é a porta de entrada por meio da qual a IA ingressa no Direito, indicam Siqueira, Morais e Santos (2022, p. 6). Na visão dos autores, a inteligência artificial precisa ser inserida num contexto de complementaridade, sob pena de seu conteúdo esvaziar-se ao longo do tempo, em concepções setorizadas de conhecimento (Siqueira; Morais; Santos, 2022, p. 6).

Segundo Hartmann Peixoto e Silva (2019, p. 21), a IA associa-se à engenhosidade humana, contribuindo com velocidade e precisão, principalmente em tarefas que demandam muito tempo e repetição de esforços. Considerado tal fluxo de ideias, é possível admitir a utilização da IA pelo Direito, diante da compatibilidade entre as tarefas exercidas pelos juristas e agentes da área e as soluções viabilizadas pela tecnologia.

A utilização de máquinas pode trazer benefícios à prática jurídica, aduzem Nunes e Marques. A implementação de ferramentas de IA na realização de pesquisas, classificação e organização de informações, busca de precedentes e elaboração de contratos tem se mostrado efetiva por proporcionar maior celeridade e precisão (Nunes; Marques, 2019, p.52).

O Laboratório de Pesquisa DR.IA, na Universidade Brasília, orienta-se pela premissa da IA como instrumento de apoio no desempenho das atividades jurídicas (Bonat; Hartmann Peixoto, 2023, p. 4). Da premissa da lógica de apoio advém o conceito de logística jurisdicional aplicada à IA, em que os desafios postos à ferramenta são instrumentais na gestão, coordenação, disposição ou organização, de modo a facilitar os desempenhos estratégicos e exercício de habilidades dos seres humanos. Essa visão orientada ao auxílio decorre do contexto de insuficiências de recursos humanos ou financeiros para fazerem frente ao desafio da ampla concretização de direitos fundamentais (Bonat; Hartmann Peixoto, 2023, p. 5-7).

Segundo Bonat e Hartmann Peixoto (2023, p. 12), a área jurídica é geradora de enorme quantidade de dados não estruturados e demandante de apoio, o que representa um bom potencial para o uso de soluções de IA. Contudo, a constante variabilidade de situações que

compõem padrões de dados jurídicos constitui um desafio a ser superado. Assim, há a necessidade de permanente formação e atualização dos sistemas de aprendizagem de máquina e o correspondente sistema de testes e avaliações.

A modernização dos sistemas demanda tempo, recursos humanos e investimentos, cujos custos são menores que soluções não tecnológicas, e resultam na melhoria de padrões de qualidade. Portando, para que haja a compatibilização de sistemas de IA com o Direito é determinante que a solução de *machine learning* seja capaz de apropriar a complexidade dos arranjos processuais jurídicos em movimentos de treinamento, testes e retreinamento (Bonat; Hartmann Peixoto, 2023, p. 12).

Com a ascensão da IA generativa (assunto que foi examinado no primeiro capítulo do presente trabalho), é possível questionar se os seus efeitos trazem algum tipo de repercussão para o Direito. Bonat e Hartmann Peixoto elaboraram metodologia para a formulação de hipóteses de impacto que associa um rol de atividades jurídicas ao provável impacto pelas famílias GPTs (Generative Pre-trained Transformer). Os autores levantam a hipótese de que atividades típicas a carreiras com padrão remuneratório maior podem ser impactadas diretamente pelo desenvolvimento das GPTs, tais como: distribuição de pessoal, homologação de atos ou edição de enunciados (Bonat; Hartmann Peixoto, 2023, p. 25- 26).

A possível influência da IA generativa na área laboral jurídica em atividades mais complexas, ou seja, para além de atividades mais simples, repetitivas e instrumentais, tem como consequência a alteração dos limites de impacto da IA. Contudo, em tarefas que apliquem conhecimentos de bases científicas ou habilidades de pensamento crítico, a exemplo da elaboração de estudos e produção de pareceres, há o decréscimo de exposição à sistemas GPTs. Em casos assim, o uso da IA é direcionado ao apoio da atividade decisória e de compreensão do Direito (Bonat; Hartmann Peixoto 2023, p. 26-28).

Nos EUA, a plataforma Harvey, que celebrou parceria com a OpenAI em 2023, tem o objetivo de fornecer produtos baseados em IA generativa para profissionais do Direito. De acordo com o site da OpenAI (2024), a Harvey fornece sistemas que auxiliam em tarefas que exigem raciocínio complexo, como elaborar documentos, responder a perguntas sobre cenários litigiosos e identificar discrepâncias materiais entre centenas de contratos. Para a pesquisa de jurisprudência, a plataforma fornece modelos personalizados que resultam da adição de 10 bilhões de tokens em dados que alimentam o modelo (OpenAI, 2024).

A utilização da plataforma Harvey[3] é restrita a escritórios de advocacia que contratem seus serviços. Contudo, os valores não são disponibilizados publicamente, de modo que os interessados em contratar devem preencher um formulário e aguardar o retorno de alguém da empresa. Segundo Stokel-Walker (2023), um dos problemas atuais das gerações de IA generativa é a sua tendência em inventar coisas com confiança - ou alucinar. Essa questão, que já é problemática nas buscas em geral, pode ser agravada na seara do Direito através do uso de sistemas como o Harvey.

Ante o exposto, contata-se que o incremento proporcionado pela IA se estende a diversas atividades relacionadas ao mercado jurídico. No presente trabalho, busca-se fazer o recorte para o uso de sistemas de IA no Poder Judiciário brasileiro, que será detalhado no próximo tópico.

2.1 Visão geral do uso da IA no Poder Judiciário brasileiro

Na tradição jurídica brasileira, no âmbito de um sistema contencioso *sui generis*, caracterizado por um amplo universo quantitativo, houve a implantação de uma série de evoluções, adaptações e soluções que se enquadram na chamada lógica analógica ou de primeiro nível. Tais soluções dependiam da atividade imediata e supervisionada dos juristas (Hartmann Peixoto; Silva, 2019, p. 51-52).

Segundo Andrade e Nunes (2023, p. 10), a doutrina tradicional apontava alternativas para as dificuldades na acessibilidade e efetividade do sistema judicial nacional, como o aumento de recursos humanos e financeiros, a extinção de certas modalidades recursais, o fomento da adoção dos chamados precedentes qualificados e ampliação dos meios alternativos de resolução de disputas (a exemplo da mediação, conciliação e arbitragem).

Constatou-se a insuficiência dos recursos de primeiro nível para o encaminhamento das demandas sociais sobre o Direito, conforme evidenciado em números e séries estatísticas (Hartmann Peixoto; Silva, 2019, p. 52). Diante disso, tem-se a contribuição decisiva da tecnologia e, especialmente, da inteligência artificial para a melhoria de um sistema altamente demandado.

O estabelecimento das bases legais para a informatização do processo judicial no Brasil ocorreu com a Lei n. 11.419/2006, que estimulou o desenvolvimento de sistemas eletrônicos para a tramitação de litígios. Apesar de terem sido criadas dezenas de plataformas para a

[3] A plataforma pode ser acessada pelo link: https://www.harvey.ai/.

tramitação eletrônica de processos, sem qualquer preocupação com a interoperabilidade, os dados digitais nelas gerados ofertam a base para tratamento e conhecimento mediante modelos de IA. Assim, sem essa primeira etapa, seria impensável o uso de inteligência artificial no Judiciário (Andrade; Nunes, 2023, p. 11).

Com o intuito de trazer maior uniformidade, a Resolução do Conselho Nacional de Justiça n. 185/2013 instituiu parâmetros para a implementação e funcionamento do Sistema Processo Judicial Eletrônico - PJe (CNJ, 2013). Posteriormente, por meio da Resolução CNJ n. 335/20, definiu-se a manutenção do PJe como sistema de processo eletrônico prioritário do CNJ; além disso, inaugurou-se política pública para a governança e a gestão de processo judicial eletrônico, de modo a buscar a integração dos tribunais do país com a criação da Plataforma Digital do Poder Judiciário Brasileiro - PDPJ-Br (CNJ, 2020). Em 2023, por meio da Portaria CNJ n. 36/23, foi instituído o Guia de Alinhamento Estratégico de Implantação da PDPJ-Br (CNJ, 2023b).

Dierle Nunes, ao apresentar a virada tecnológica do direito processual, propõe a adaptação procedimental mediante o emprego de tecnologia, a exemplo da automação de atos e fatos processuais, ODRs e o emprego de inteligência artificial. Desse modo, os processos seriam adaptados tecnologicamente com ferramentas de auxílio nas atividades processuais, criando-se novas vias mais adequadas de dimensionamento de conflitos (Nunes, 2023, p. 415-416).

A virada tecnológica indicada por Nunes não é apenas a organização do fluxo de trabalho (*workflow*) que aumenta a eficiência e diminui o tempo de um processo eletrônico, pois segundo o autor tarefas como a estruturação das etapas e a redução dos tempos mortos de juntada podem ser automatizadas sem qualquer emprego de IA. Por meio da inteligência artificial, é possível parametrizar uma grande massa de dados (*big data*), tratando informações que se encontram desestruturadas e com isso se obter uma revolução nos institutos jurídicos (Nunes, 2023, p. 418).

O Programa Justiça 4.0, que é fruto de parceria entre o Conselho Nacional de Justiça (CNJ) e o Programa das Nações Unidas para o Desenvolvimento (PNUD), com apoio do Conselho da Justiça Federal (CJF), Superior Tribunal de Justiça (STJ), Tribunal Superior do Trabalho (TST) e Conselho Superior da Justiça do Trabalho (CSJT), busca tornar “o sistema judiciário brasileiro mais próximo da sociedade ao disponibilizar novas tecnologias e inteligência artificial” e promover “soluções digitais colaborativas que automatizam as

atividades dos tribunais, otimiza o trabalho dos magistrados, servidores e advogados" (CNJ, [s.d]).

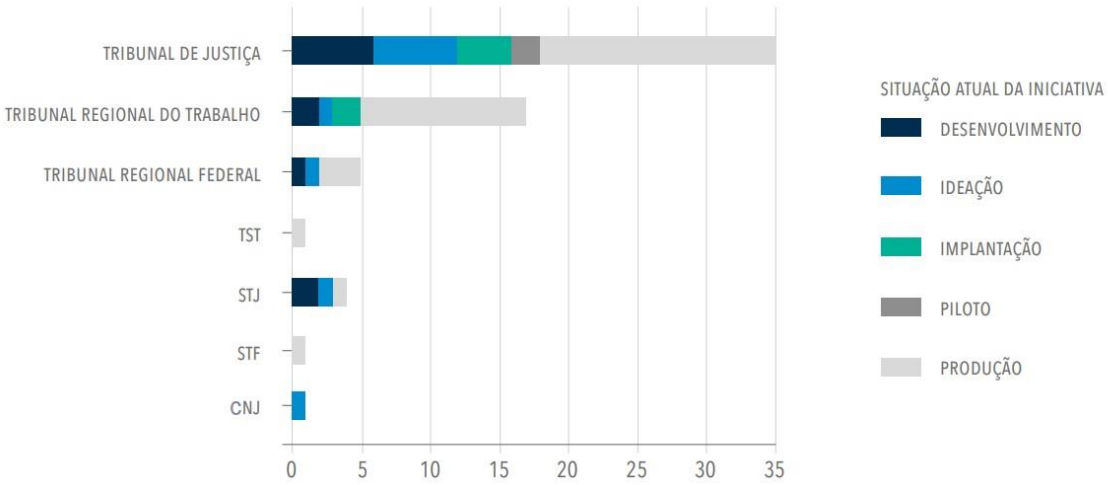
A proposta do Programa Justiça 4.0 em desenvolver ações, estudos e estratégias para ampliar a prestação jurisdicional e facilitar o acesso à justiça engloba o Juízo 100% Digital; o Balcão Virtual; a Plataforma Digital do Poder Judiciário (PDPJ); o auxílio aos tribunais nos registros processuais primários; a consolidação, implantação, tutoria, treinamento, higienização e publicização da Base de Dados Processuais do Poder Judiciário (DataJud); a plataforma Codex e a Plataforma Sinapses (Bragança; Loss; Braga, 2022, p. 41-42),

A Plataforma Sinapses foi instituída através da Resolução CNJ n. 332/2020 como uma plataforma nacional de armazenamento, treinamento supervisionado, controle de versionamento, distribuição e auditoria dos modelos de inteligência artificial. No âmbito do Sinapses é disponibilizado o Repositório Nacional de Projetos de Software e Versionamento de Arquivos (Git.Jus) para utilização pelos tribunais, permitindo o compartilhamento de soluções computacionais construídas para interação com o PJe (CNJ, 2021).

Na primeira edição da pesquisa "Tecnologia Aplicada à Gestão dos Conflitos no âmbito do Poder Judiciário Brasileiro", sob a coordenação do Ministro Luis Felipe Salomão, foi apontado que, de forma geral, os projetos de IA nos tribunais comportam as seguintes funcionalidades: verificação das hipóteses de improcedência liminar do pedido nos moldes enumerados pelos incisos do art. 332 do CPC; sugestão de minuta; agrupamento de processos por similaridade; realização do juízo de admissibilidade dos recursos; classificação dos processos por assunto; penhora online; extração de dados de acórdãos; reconhecimento facial; *chatbot*; classificação de petições; indicação de prescrição; padronização de documentos; transcrição de audiências; distribuição automatizada e classificação de sentenças (Salomão, 2020, p. 69).

Na segunda edição da mencionada pesquisa, publicada em 2022, quarenta e quatro tribunais responderam a um questionário que reuniu diversos aspectos práticos da utilização e desenvolvimento de iniciativas de IA. A partir disso, realizou-se um mapeamento das fases de implantação, em que se constatou que os tribunais de justiça possuem uma relevante distribuição de iniciativas nas diferentes fases de adoção e desenvolvimento, enquanto outros tribunais possuem poucas iniciativas, mas que já estão em produção, o que mostra um nível de maturidade maior (Salomão, 2022, p. 252; p. 255). O gráfico abaixo reproduzido detalha esse contexto:

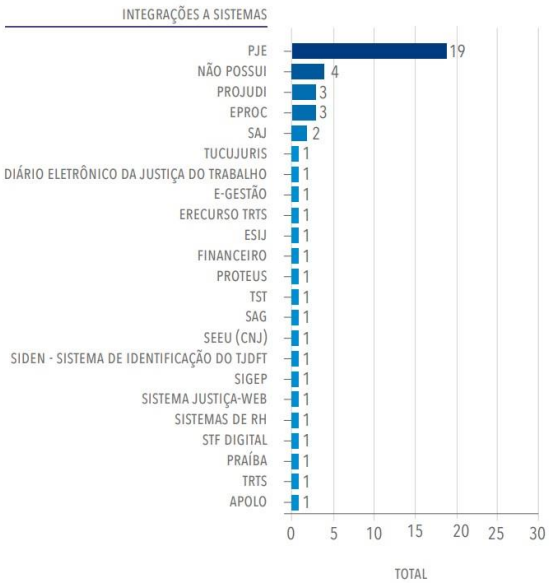
Figura 9 - Situação das iniciativas de IA por tribunal



Fonte: Salomão, 2022.

Através de mecanismos de IA, é possível aprimorar funcionalidades de sistemas de processo judicial eletrônico (Siqueira; Morais; Santos, 2022, p. 3). Nessa lógica, o relatório da pesquisa "Tecnologia Aplicada à Gestão dos Conflitos no âmbito do Poder Judiciário Brasileiro" identificou que o PJe estava integrado a 19 iniciativas de utilização de serviços inteligentes e apenas quatro dos projetos investigados não tinham integração com outros sistemas dos tribunais (Salomão, 2022, p. 256; p. 257).

Figura 10 - Integração das iniciativas de IA a sistemas eletrônicos



Fonte: Salomão, 2022.

Existem projetos de implantação de ferramentas de IA em todos os Tribunais Superiores e em todos ramos da justiça: Justiça Estadual, Justiça Federal, Justiça do Trabalho, Justiça Eleitoral e Justiça Militar (CNJ, 2023). Convém ressaltar que a primeira iniciativa do tipo foi o Projeto Victor no Supremo Tribunal Federal, em 2018, época em que poucos no Direito estavam familiarizados com os alcances da IA (Hartmann Peixoto, 2020b, p. 2).

Realizado integralmente pela Universidade de Brasília (UnB), com o apoio e a participação dos servidores do STF, o Projeto Victor teve como objetivo aplicar métodos de aprendizado de máquina para o reconhecimento de padrões nos processos jurídicos relativos a julgamentos de repercussão geral na Corte Constitucional. Em resumo, a IA do Victor busca analisar texto do processo para classificá-lo em algum tema reconhecido de repercussão geral (Hartmann Peixoto, 2020b, p. 3).

O Projeto Victor foi a vanguarda de um movimento de inovação no Judiciário que tem como corolário inúmeros outros projetos. Segundo Salomão e Tauk (2023b, p. 11), os modelos computacionais de IA no Judiciário podem ser divididos, para fins didáticos, em quatro grupos principais em relação às tarefas que desempenham. No primeiro grupo, inclui-se os sistemas relacionados às atividades- meio, que objetivam gerir de modo mais eficiente os recursos financeiros e de pessoal, a exemplo do *Chatbot* Digep, do Tribunal de Justiça do Rio Grande do Sul (TJRS), que responde dúvidas dos servidores quanto aos assuntos relacionados à gestão de pessoas.

No segundo grupo, que contempla ferramentas de IA relacionadas à atividade-fim do Judiciário, está situada a maioria dos modelos, os quais se destinam à automação dos fluxos de movimentação dos processos e o apoio aos juízes em tarefas pré-determinadas. São exemplos desse grupo, o Athos, no Superior Tribunal de Justiça, que faz a identificação e o monitoramento de temas repetitivos e o Mandamus, no Tribunal de Justiça de Roraima - TJRR (Salomão; Tauk, 2023b, p. 11-12). Com o objetivo de aprimorar as etapas de citação e intimação, o Mandamus foi desenvolvido para apoiar a execução de mandados na identificação, estruturação de formatos e distribuição das ordens, otimizando o tempo de cumprimento da diligência e permitindo o aperfeiçoamento dos recursos humanos envolvidos (Bonat; Hartmann Peixoto, 2021, p. 10-11).

Salomão e Tauk (2023b, p. 12) inserem no terceiro grupo as soluções baseadas em IA que colaboram na elaboração de minutas com conteúdo decisório de sentença, votos ou decisões interlocutórias. Entre as ferramentas representativas do grupo está o projeto Alei, no Tribunal

Regional Federal da 1ª Região (TRF1), que dentre outras tarefas, busca associar ao processo judicial em análise julgados anteriores e jurisprudências.

Por fim, no quarto grupo situam-se iniciativas relacionadas a formas adequadas de resolução de conflitos, em que a partir de informações de processos similares tem-se o amparo às partes na busca da melhor solução. O Índice de Conciliabilidade por Inteligência Artificial (ICIA), no Tribunal Regional do Trabalho da 4ª Região (TRT4), é citado como exemplo desse grupo, já que estima a probabilidade de um processo ser conciliado no estágio em que se encontra (Salomão; Tauk, 2023b, p. 12-13).

Correa e Gonçalves (2022, p. 295-296) apontam que no Brasil, a busca da solução de um conflito assume predominantemente natureza contenciosa, em que há o intuito de obter um documento revestido do poder coercitivo do Estado, o que se consubstancia na chamada "cultura da sentença". A insuficiência do Poder Judiciário em oferecer um resultado célere e útil fez despontar a gestão racional dos conflitos por meio da "cultura do consenso", em que as próprias partes interessadas, com ou sem a colaboração de um terceiro, buscam resolver o conflito (Correa; Gonçalves, 2022, p. 300). Assim, sistemas de IA voltados à autocomposição podem ser formas de estímulo a essa cultura, ainda que inseridas no próprio judiciário.

Segundo o Painel de Projetos de IA no Poder Judiciário, disponibilizado pelo CNJ e com dados atualizados até maio de 2022, havia 111 projetos de IA em uso e/ou em desenvolvimento no Judiciário brasileiro, um aumento de 171% em relação ao ano de 2021. Dos 88 tribunais que participaram da pesquisa, 53 contavam com projetos de IA. Além disso, estavam inseridos no Sinapses 42 projetos. Ao responderem à pergunta sobre a motivação para o uso de ferramentas de IA, as respostas mais recorrentes dadas pelos tribunais foram: aumento de produtividade (volume/tempo), inovação, melhoria da qualidade dos serviços, redução de custos (CNJ, 2022).

Na atualização de 2023, o Painel de Projetos de IA no Poder Judiciário apresentou um total de 140 projetos de IA em uso e/ou em desenvolvimento no Judiciário brasileiro, o que representa aumento de 26% em relação ao ano de 2022. Dos 94 tribunais que participaram da pesquisa, 62 contavam com projetos de IA. Acerca dos benefícios alcançados com a implementação dos projetos, as respostas mais frequentes fornecidas pelos tribunais foram: maior eficiência e agilidade no processamento de documentos e informações; otimização de recursos e redução de custos operacionais; automatização de tarefas repetitivas e burocráticas; redução do tempo de tramitação dos processos. Contudo, o maior desafio para implementação,

segundo os tribunais, é a dificuldade em encontrar profissionais qualificados para trabalhar com inteligência artificial (CNJ, 2023).

O ano de 2023 se encerrou com um acervo de 83,8 milhões de processos em tramitação, os quais 90,6% eram eletrônicos, segundo dados do relatório Justiça em Números. Nesse mesmo ano foram recebidos 3 milhões de casos novos a mais do que em 2022, mas a alta produtividade atenuou esse impacto, que não foi totalmente refletido no saldo de elevação do acervo (CNJ, 2024, p. 15; p.28). Desse modo, a utilização da IA tem uma importância significativa no contexto brasileiro, dada a judicialização expressiva e a necessidade de racionalizar recursos (Salomão; Tauk, 2023, p. 8).

Os algoritmos já estão em uso em muitos domínios e desempenham papeis cada vez maiores, porém, é pouco provável que estes venham a substituir o julgamento humano no estágio final de decisões importantes, afirmam Kahneman, Sibony e Sunstein (2022, p. 359). Desse modo, os autores consideram que apesar de inviável a substituição do julgamento humano, este pode ser melhorado através do uso da IA.

Dispor de sistemas baseados em IA no Poder Judiciário não é sinônimo de delegar à máquina o poder de julgamento, até porque esse não é o propósito das atuais ferramentas em uso pelos tribunais brasileiros. O que se espera é o uso ferramentas/soluções para aprimorar o trâmite processual e aperfeiçoar a tomada de decisões pelos magistrados, na medida em que a IA auxilia a checagem de dados, viabiliza a indexação de informações, permite a busca de jurisprudência mediante pesquisas semânticas e fornece um poderoso auxílio em tarefas repetitivas (Andrade; Nunes, 2023, p. 23).

Como órgão vocacionado à resolução de conflitos de interesses e tutela de direitos fundamentais, o Poder Judiciário é constantemente chamado a intervir em questões sensíveis envolvendo a vida em sociedade. A abertura do sistema jurídico para as soluções baseadas em IA é capaz de minimizar o movimento expansivo de aumento de acervos processuais ao longo dos últimos anos, a mora excessiva no curso do processo e a deflagrar redução de custos com a máquina judiciária (Siqueira; Morais; Santos, 2022, p. 25).

Segundo Fernanda de Carvalho Lage (2021, p. 143), o uso da IA pelo campo jurídico é altamente promissor, mas é essencial que haja o compartilhamento de experiências pelos tribunais para que sejam reduzidos custos e esforços repetitivos pelas Cortes. Ademais, considerando que o universo das abordagens baseadas em inteligência artificial é vasto e diverso em termos das tecnologias e problemas atacados, à medida que tais iniciativas ganharem

mais força, aumenta a necessidade de investimento em infraestrutura adequada ao processamento de alto desempenho dentro do Judiciário (Salomão, 2022, p. 265).

Por fim, é preciso refletir se as ferramentas baseadas em IA utilizadas pelo Judiciário exigem alguma ação tecnológica do jurisdicionado. Nesse sentido, Nunes (2023, p. 427) aponta que diante da incorporação da tecnologia ao direito processual, não se pode negligenciar os dilemas relacionados à carência de cidadania digital de parcela da população, que decorre da falta de equipamento, letramento e conexão de qualidade, de tal maneira que se reivindique uma justiça híbrida, meio presencial e meio digital.

2.2 O papel da regulação na construção e no uso responsável de sistemas de IA

Segundo Marçal Justen Filho (2023, p. 511), a regulação econômico-social consiste na "atividade estatal de intervenção indireta sobre a conduta dos sujeitos públicos e privados, de modo permanente e sistemático, para implementar as políticas de governo e a realização dos direitos fundamentais". Para o autor, a expressão "regulamentação" corresponde ao desempenho de função normativa infraordenada, pela qual se detalham as condições de aplicação de uma norma de cunho abstrato e geral. Apesar da regulação ser um conceito mais amplo, eventualmente esta pode resultar em atos de regulamentação.

A regulação é um dos tipos de atividade estatal que se traduz no desempenho tanto de função administrativa como legislativa, jurisdicional e de controle. Além disso, essa atuação não é considerada um fim em si mesma, mas um instrumento para promover conscientemente os fins essenciais do Estado (Justen Filho, 2023, p. 512). Desse modo, refletir sobre a regulação da inteligência artificial é sobretudo compreender a atuação estatal frente a um fenômeno global que vem resultando em mudanças de paradigmas.

Segundo Russell (2022, p. 47), é necessário compreender as aplicações potenciais da IA não apenas como problemas tecnológicos a serem resolvidos, mas também a sua inserção num contexto social. O sucesso deve ser medido não pela precisão das previsões e decisões do sistema, mas pelas consequências reais da sua implantação. O autor defende a necessidade de uma teoria de integração sociotécnica para sistemas de IA, algo análogo ao papel que o planejamento urbano desempenha para os artefatos produzidos pela engenharia civil e pela arquitetura.

Um mesmo algoritmo, quando utilizado para uma finalidade ou contexto diversos poderá alterar sua relevância social, e assim conduzir a diferentes necessidades regulamentares.

O reconhecimento facial usado para desbloquear um telefone não é tão problemático como quando aplicado à vigilância em massa. Portanto, sendo a IA um fenômeno heterogêneo, a sua regulação não pode ser única unitária (Bertolini, 2020, p. 9).

Para que se possa pensar em uma regulação adequada é imperioso entender a sua exequibilidade por parte daqueles que produzem e usam ferramentas de IA, de forma a questionar o espectro de liberdade de aplicação das inovações tecnológicas e também a responsabilidade pelas consequências ante a ocorrência de certos tipos de danos, destacam Lage e Hartmann Peixoto (2021, p. 279). Ademais, os autores acordam que a regulação não deve reprimir os importantes avanços científicos.

A regulação excessiva da IA pode ter um efeito inibidor na inovação, enquanto a subregulação pode ter sérios impactos nos direitos dos cidadãos. Assim, a forma como lidamos com o assunto definirá o mundo em que vamos viver. Uma abordagem regulatória ampla, que tentasse incluir todas as utilizações existentes da IA estaria condenada a ser incompleta e ineficaz, porque, por ser demasiadamente genérica, pode não focar adequadamente nas peculiaridades que dão origem a preocupações e oportunidades relevantes para a sociedade (Bertolini, 2020, p. 21-32).

O Estado e o Direito devem extrair da IA seus incontestáveis benefícios, mas não podem se furtar de regular o seu uso e desenvolvimento à luz da ética e dos princípios insculpidos na Constituição, defendem Andrade e Nunes (2023, p. 22). Para os autores, enxergar a inteligência artificial à luz da ética e dos direitos fundamentais é uma janela de oportunidade para moldar seu desenvolvimento, de forma a garantir a sua evolução de maneira segura, oferecendo um ambiente de operação confiável aos usuários (Andrade; Nunes, 2023, p. 22).

A abordagem adequada dos riscos das aplicações de IA envolverá inevitavelmente a adaptação dos sistemas jurídicos para responderem melhor ao rápido avanço do ritmo do desenvolvimento tecnológico, pois enquanto a IA avança velozmente, os governos tendem a adaptar-se aos novos desafios de forma mais morosa. Os esforços de experimentação e teste são um aspecto importante para a elaboração de regras nesse campo e podem ocorrer tanto em ambientes do mundo real quanto em ambientes simulados (Littman et al., 2021, p. 42).

2.3 Perspectivas da regulação da IA no Brasil

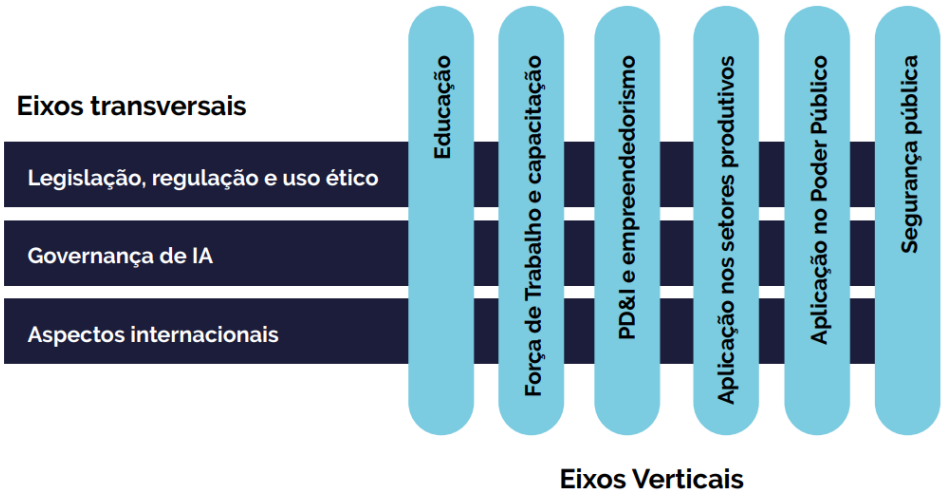
No âmbito da Estratégia Brasileira para a Transformação Digital (E-Digital), cuja estrutura de governança para implantação foi estabelecida pelo Decreto n. 9.319/18, já havia a

sinalização da importância das novas tecnologias digitais e da proteção de direitos no ambiente digital. O referido documento indica que "É preciso avaliar as implicações jurídicas e éticas de aplicações de inteligência artificial, Internet das Coisas e outras áreas da fronteira tecnológica" (E-Digital, 2018, p. 41).

Em abril de 2021, a Portaria do Ministério da Ciência Tecnologia e Inovação n. 4.617 instituiu a Estratégia Brasileira de Inteligência Artificial - EBIA e seus eixos temáticos. De acordo esse ato normativo secundário, a EBIA tem como finalidades "nortear as ações do Estado brasileiro em prol do fortalecimento da pesquisa, desenvolvimento e inovações de soluções em Inteligência Artificial, bem como, seu uso consciente, ético para um futuro melhor"; e "garantir a inovação no ambiente produtivo e social na área de Inteligência Artificial, capaz de enfrentar os desafios associados ao desenvolvimento do País" (Brasil, 2021).

Pontua-se que o documento de referência da EBIA, com base nas diretrizes para a inteligência artificial da OCDE, a fim de organizar o debate acerca de uma Estratégia Brasileira de IA, estabeleceu 9 (nove) eixos temáticos (EBIA, 2021, p. 7), abaixo identificados:

Figura 11 - Eixos temáticos da Estratégia Brasileira de IA



Fonte: EBIA, 2021.

A EBIA é pautada por 6 (seis) objetivos estratégicos que levam em consideração o ecossistema tecnológico, são eles: i) Contribuir para a elaboração de princípios éticos para o desenvolvimento e uso de IA responsáveis[4]; ii) Promover investimentos sustentados em

[4] Os princípios éticos indicados pela EBIA serão detalhados nos tópicos 3.3 e 3.4 do trabalho.

Pesquisa e Desenvolvimento em IA; iii) Remover barreiras à inovação em IA; vi) Capacitar e formar profissionais para o ecossistema da IA; v) Estimular a inovação e o desenvolvimento da IA brasileira em ambiente internacional; vi) Promover ambiente de cooperação entre os entes públicos e privados, a indústria e os centros de pesquisas para o desenvolvimento da Inteligência Artificial (EBIA, 2021, p. 8).

Segundo a EBIA, no centro dos debates acerca do desenvolvimento de parâmetros jurídicos, regulatórios e éticos que orientam o uso da IA, encontra-se a preocupação em estabelecer um ponto de equilíbrio entre a proteção e a salvaguarda de direitos; a preservação de estruturas adequadas à evolução de uma tecnologia cujas potencialidades ainda não foram plenamente compreendias; e a garantia de segurança jurídica quanto à responsabilização dos diferentes atores que participam da cadeia de valor de sistemas autônomos (EBIA, 2021, p. 17).

No tocante à promoção de um ambiente institucional e regulatório propícios à inovação e ao desenvolvimento tecnológico, a Estratégia Brasileira de IA posiciona-se no sentido de que diante de um cenário de rápida evolução, a normatização é complexa e propensa a se tornar obsoleta rapidamente. Assim, o documento preconiza que antes de se adotar novas leis, regulações ou controles, os governos possam avaliar e refletir sobre a conjuntura (EBIA, 2021, p. 18).

A Lei Geral de Proteção de Dados Pessoais - LGPD (Lei n. 13.709/18) tornou-se fundamental na inserção do Brasil no rol de países que possuem uma legislação voltada à proteção de dados pessoais, repercutindo inclusive na Emenda Constitucional n. 115 de 2022, que incluiu a referida proteção no rol dos direitos fundamentais do art. 5º da CF/88. A LGPD é necessária para garantir o desenvolvimento de sistemas de IA de confiança, que são baseados no processamento e na transferência de dados pessoais mediante o consentimento livre e informado (Oliveira, 2022, p. 147).

Além da LGPD, que é uma norma que se relaciona de maneira direta com a IA, a EBIA (2021, p. 20) expõe que o Decreto n. 8.777/2016, que institui a Política de Dados Abertos do Poder Executivo Federal também se conecta com essa tecnologia, já que bases de dados abertos podem servir para a alimentação de sistemas, o que destaca a importância de diretrizes sobre o uso ético de dados abertos. Na Estratégia, também se evidencia a relevância do compartilhamento de soluções de software entre as esferas de governo, o qual é respaldado pela Portaria STI/MP n. 46/2016 e pela lei n. 14.063/20.

Para Gaspar e Mendonça, a Estratégia Brasileira de Inteligência Artificial (EBIA) não indica questões de planejamento que seriam básicas para a implementação de uma estratégia bem-sucedida, fazendo com que o documento se assemelhe a uma carta de intenções. Aponta-se que há uma indefinição da estrutura de governança, caracterizada pela não identificação de atores responsáveis e respectivas capacidades na concretização das ações; não há um delineamento acerca de prazos e metas; e, a generalidade excessiva faz com que as questões levantadas não sejam devidamente debatidas (Gaspar; Mendonça, 2021).

Outro ponto objeto de crítica é a inexistência de dotação orçamentária e recursos para o desenvolvimento dos eixos da EBIA. Portanto, é difícil vislumbrar de que modo os objetivos serão realizados ao longo do tempo. Além disso, o documento contém a proposta de que a regulação dos sistemas de IA ocorrerá com a cautela de não interferir na inovação tecnológica. Contudo, como não são apresentados parâmetros para que essa receita seja seguida, assim, caberá ao Poder Judiciário decidir quando os conflitos relativos ao tema surgirem (Oliveira, 2022, p. 149-150).

A Resolução CNJ n. 332/2020 é um ato normativo que "dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário e dá outras providências". A Resolução dedica-se aos seguintes pontos ao longo de seus capítulos: aspectos gerais; respeito aos direitos fundamentais; não discriminação; publicidade e transparência; governança e qualidade; segurança; controle do usuário; pesquisa, desenvolvimento e implantação de serviços de inteligência artificial; prestação de contas e responsabilização (CNJ, 2020b).

Segundo Salomão e Tauk (2023b, p. 16), a Resolução CNJ n. 332/2020 foi inspirada na Carta Europeia de Ética sobre o Uso da Inteligência Artificial em Sistemas Judiciais e seu ambiente, de 2018, de modo que ambos os documentos adotam princípios éticos similares. Os cinco princípios fundamentais adotados pela Carta Europeia são: respeito aos direitos fundamentais; não-discriminação; qualidade e segurança; transparência, imparcialidade e equidade; e, controle do usuário (CEPEJ, 2018), o quais também são detalhados no referencial brasileiro.

Evidencia-se que a Resolução CNJ n. 332/2020 dispõe que "A utilização de modelos de Inteligência Artificial em matéria penal não deve ser estimulada, sobretudo com relação à sugestão de modelos de decisão preditivas", ressalvando a utilização de soluções computacionais destinadas à automação e ao oferecimento de subsídios destinados ao cálculo

de penas, prescrição, verificação de reincidência, mapeamento, classificações e triagem. De acordo com Hartmann Peixoto (2020, p. 50), o texto da Resolução expressa atenção ao considerar as experiências internacionais negativas de preconceitos no uso/desenvolvimento de IA em matéria penal.

Nota-se ainda que a Portaria CNJ n. 271/2020, cuja ementa dispõe que esta "regulamenta o uso de Inteligência Artificial no âmbito do Poder Judiciário", tem disposições que complementam a Resolução CNJ n. 332/2020. Na Portaria consta, entre outras matérias, o que são considerados projeto de IA voltados ao Judiciário, os requisitos a serem observados na Pesquisa e Desenvolvimento de soluções nesse âmbito e reforça a utilização do Sinapses como plataforma comum (CNJ, 2020c).

Tauk e Navarro (2022, p. 49) destacam que o CNJ vem se empenhando para manter o Poder Judiciário em sintonia com as exigências sociais, jurídicas e éticas para o uso de tecnologia e dos sistemas de IA. Nessa perspectiva, a Portaria CNJ n. 338/2023 instituiu Grupo de Trabalho sobre Inteligência Artificial no Poder Judiciário, cujo objetivo é realizar estudos e apresentar proposta de regulamentação do uso de sistemas de IA generativa baseada em grandes modelos de linguagem no Poder Judiciário. O mencionado grupo é composto por trinta membros e tem previsão de encerramento de suas atividades em um ano, a contar da publicação da Portaria (CNJ, 2023c).

Um outro ato tendente a estabelecer disposições relacionadas ao uso de inteligência artificial, porém circunscrita ao âmbito eleitoral, é a Resolução n. 23.610/2019, do Tribunal Superior Eleitoral. A mencionada Resolução, a qual dispõe sobre propaganda eleitoral, foi alterada pela Resolução 23.732/24 e passou a estabelecer que a utilização, na propaganda eleitoral, de conteúdo sintético multimídia gerado por meio de inteligência artificial para criar, substituir, omitir, mesclar ou alterar a velocidade ou sobrepor imagens ou sons impõe ao responsável o dever de informar, de modo explícito, destacado e acessível que o conteúdo foi fabricado ou manipulado e a tecnologia utilizada (TSE, 2024).

As atualizações que agora estão presentes na Resolução TSE n. 23.610/2019 e que se associam à IA incluem também: proibição das *deepfakes*; restrição do emprego de robôs para intermediar contato com o eleitor, sendo proibida a simulação de diálogo com o candidato ou outra pessoa real; e responsabilização das *big techs* que não retirarem do ar, imediatamente, conteúdos com desinformação, discurso de ódio, além dos antidemocráticos, racistas e homofóbicos (TSE, 2024b).

2.4 Inteligência Artificial em pauta no Poder Legislativo

Segundo Cristina Godoy Bernardo de Oliveira (2022, p. 153), é relevante compreender os projetos de lei propostos no âmbito federal que buscam regular juridicamente a IA, a fim de se analisar quais serão os possíveis caminhos a serem seguidos pelo Brasil.

A partir do ano de 2019, notou-se a movimentação de deputados e senadores quanto à apresentação de Projetos de Lei em suas respectivas Casas com propostas que têm como centro de discussão a inteligência artificial no país. Segundo Oliveira (2022, p. 155), até 2022, o projeto mais debatido e com maior chance de aprovação era o PL n. 21/2020.

Apresentado na Câmara dos Deputados pelo deputado federal Eduardo Bismarck, o PL n. 21/2020 foi aprovado com alterações pelo Plenário da casa iniciadora em 29/09/2021, sendo então encaminhado ao Senado Federal. Segundo a ementa, o referido projeto "Estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil; e dá outras providências" (Brasil, 2021).

Os princípios estabelecidos no PL n. 21/2020 são semelhantes ao arcabouço principiológico da LGPD, aponta Oliveira (2022, p. 155), o que demonstra certa redundância. Além disso, há o risco de ser aprovada uma lei vaga, sem as devidas definições, taxonomia e classificação de risco para que seja considerado um Marco Legal de IA, que permitiria uma regulação setorial posterior (Oliveira, 2022, p. 155).

No contexto de discussões acerca do Substitutivo ao PL n. 21/20, foi apresentada nota de colaboração pelo Laboratório de Pesquisa DR.IA, da UnB. Na manifestação, é apontado que a proposta legislativa apresenta sérios problemas que comprometeriam o desenvolvimento tecnológico do país na exploração robusta e sustentável de dados por meio da IA. Entre as considerações transmitidas, destaca-se a sugestão de inclusão das universidades como um dos três componentes internacionalmente considerados para o desenvolvimento da IA, juntamente com o poder público e o setor produtivo (Hartmann Peixoto, 2021, p. 1; p. 4).

Outro ponto de atenção, divulgado na mencionada nota de colaboração, refere-se ao princípio da transparência. É exposto que o Substitutivo retira o caráter geral e dinâmico desse princípio, que deve ter uma previsão que integre a recorrente opacidade das arquiteturas algorítmicas dos sistemas de IA. Assim, diante da complexidade para a concretização do princípio da transparência, é sugerida uma nova redação de caráter mais amplo que ameniza o problema da proposta legislativa (Hartmann Peixoto, 2021, p. 5).

Em nota técnica sobre o Substitutivo ao Projeto de Lei n. 21/20 enviado ao Senado Federal, elaborada pelo Centro de Inovação, Administração e Pesquisa do Judiciário da FGV (CIAPJ FGV), é reconhecido que este é um verdadeiro marco legal para o desenvolvimento e uso da IA no país, pois uma vez aprovada a lei, terá os atributos da abstratividade, generalidade, imperatividade e coercibilidade. Contudo, a nota ressalta que quanto à sua eficácia normativa, o texto não prevê claramente as responsabilidades e as penalidades para os casos de descumprimento (Salomão, 2022b, p. 12-13).

De acordo com o parecer do CIAPJ FGV, a previsão de um único regime de responsabilidade, em relação às ações danosas dos sistemas de inteligência artificial (subjetiva, salvo previsão legal), desconsidera a diversidade de sistemas, a multiplicidade de agentes envolvidos e a variedade de relações jurídicas que podem ocorrer. Os sistemas de inteligência artificial são diferentes, operam em setores distintos e geram riscos diversos, portanto, não é uma área homogênea (Salomão, 2022b, p. 20).

Proposto em 2023, o PL n. 2.338, de autoria do senador Rodrigo Pacheco, tem como propósito estabelecer normas gerais de caráter nacional para o desenvolvimento, implementação e uso responsável de sistemas de IA. Na justificativa da proposição, afirma-se que esta tem um duplo objetivo, pois "estabelece direitos para proteção do elo mais vulnerável em questão, a pessoa natural que já é diariamente impactada por sistemas de inteligência artificial"; e, por outro lado, dispõe de "ferramentas de governança e de um arranjo institucional de fiscalização e supervisão", que cria condições de segurança jurídica para a inovação e o desenvolvimento tecnológico (Brasil, 2023).

Destaca-se que o PL n. 2.338/23 destinou uma seção específica para o tema da Avaliação de Impacto Algorítmico e indica que esta é uma "obrigação dos agentes de inteligência artificial, sempre que o sistema for considerado como de alto risco pela avaliação preliminar". Os agentes de inteligência artificial, segundo a proposta, são os fornecedores e operadores de sistemas de inteligência artificial; e a citada avaliação preliminar consiste em uma classificação do grau de risco de um sistema de IA, realizada pelo fornecedor (Brasil, 2023).

Acioly, Mendes e Monteiro Neto (2023, p. 246), ao analisarem a mencionada proposta legislativa, reputam que a funcionalidade das avaliações de impacto demonstra compatibilidade com a necessidade de explicabilidade dos sistemas de inteligência artificial, considerando-se os riscos técnico-jurídicos que são consignados pela opacidade dos algoritmos de aprendizagem. Para os autores, a estratificação de riscos conduz a uma carga de obrigações proporcional ao

risco atribuído no caso concreto, impondo-se uma estrutura de governança que propicie proteção dos direitos fundamentais dos possíveis afetados por esse tipo ferramenta (Acioly; Mendes; Monteiro Neto, 2023, p. 246).

Os Projetos de Lei n. 145/2024 e n. 146/2024, de iniciativa do senador Chico Rodrigues (Brasil, 2024b), trazem questões pontuais sobre uso de sistemas de inteligência artificial para manipulação, processamento e geração de imagens, sons, áudios e vídeos nas áreas do Direto do Consumidor e Direito Penal.

A proposta de alteração do Código de Defesa do Consumidor (Lei n. 8.078/90) contida no PL n. 145/24 tem o intuito de "regular o uso de ferramentas de inteligência artificial para fins publicitários e coibir a publicidade enganosa com o uso dessas ferramentas". Objetiva-se acrescentar o art. 37-A no CDC, para proibir a publicidade "em que a imagem ou voz de pessoa, viva ou falecida, seja manipulada mediante o emprego de sistemas de inteligência artificial para o processamento, análise e geração de imagens e áudio com o intuito de influenciar a percepção do consumidor quanto ao produto ou serviço e promover sua comercialização", ressalvados os casos de consentimento do titular ou a prestação de informações ostensivas ao consumidor de que se trata de publicidade elaborada mediante o uso de IA (Brasil, 2024b).

O Projeto de Lei n. 146/2024, por sua vez, tem como finalidade estabelecer "causa de aumento de pena para os crimes contra a honra e hipótese qualificada para o crime de falsa identidade, para quando houver a utilização de tecnologia de inteligência artificial para alterar a imagem de pessoa ou de som humano". Assim, consoante as disposições do PL, nos crimes de calúnia, injúria e difamação seriam acrescentadas duas hipóteses de aumento de pena; e no crime de falsa identidade seria estabelecida a pena de um a cinco anos, e multa, para a hipótese qualificada pelo uso de IA. (Brasil, 2024c).

Partindo para uma análise mais principiológica, o PL n. 210/2024, de autoria do senador Marcos do Val, dispõe "sobre os princípios para uso da tecnologia de inteligência artificial no Brasil". De modo similar ao PL n. 21/2020, que traz definição para "sistema de inteligência artificial" (Brasil, 2021), a proposta citada busca conceituar o que são "tecnologias de inteligência artificial". O Projeto possui nove artigos e enumera apenas cinco princípios que devem ser observados no uso de IA; são eles: segurança e efetividade dos sistemas; proteção contra discriminação de algoritmo; garantia à privacidade dos dados e informações; direito à informação; e direito à opção pelo tratamento humano e direito à contestação (Brasil, 2024d).

Não é apresentada na justificativa do PL os critérios e parâmetros utilizados para a escolha de tais princípios.

O Projeto de Lei n. 266/2024, apresentado pelo senador Veneziano Vital do Rêgo, dispõe "sobre o uso de sistemas de inteligência artificial para auxiliar a atuação de médicos, advogados e juízes". O escopo da proposta é essencialmente alterar normas relacionadas a essas três carreiras, através de acréscimos na Lei n. 12.842/13 (dispõe sobre o exercício da medicina), na Lei n. 8.906/94 (dispõe sobre o estatuto da advocacia e a OAB), no Decreto-Lei n. 2.848/40 (Código Penal) e na Lei n. 13.105/15 (Código de Processo Civil). No CPC, por exemplo, propõe-se o acréscimo do art. 194-A para dispor que "Sistemas de inteligência artificial poderão ser utilizados para auxiliar a prática de atos processuais". Na justificativa, argumenta-se que nas carreiras abrangidas pela proposição o mau uso da tecnologia pode representar um alto risco para a sociedade (Brasil, 2024e).

No âmbito do Senado Federal, a presidência determinou em fevereiro de 2024, com fulcro no art. 48, § 1º, do Regimento Interno, a tramitação conjunta dos Projetos de Lei n. 5.051 e n. 5.691, de 2019; n. 21, de 2020; n. 872, de 2021; n. 2.338 e n. 3.592, de 2023; e n. 145, n. 146, n. 210 e n. 266, de 2024, por tratarem de tema correlato. Determinou-se que "As matérias passam a tramitar em conjunto e vão ao exame da Comissão Temporária sobre Inteligência Artificial no Brasil - CTIA" (Brasil, 2024). Além das proposições mencionadas, identificou-se na página eletrônica do Senado Federal[5] outros nove Projetos de Lei que tramitam na Casa e estão diretamente relacionados à inteligência artificial, são os Projetos n. 1.197, n. 1.833, n. 262, n. 370 e n. 2.024, de 2024; n. 5.721, n. 5.722, n. 6.065 e n. 1.272, de 2023 (Brasil, 2024f).

No endereço eletrônico da Câmara dos Deputados, ao ser realizada a busca pelos Projetos de Lei em tramitação cujo assunto é "inteligência artificial"[6], o site da Casa Legislativa traz o resultado de que cento e doze proposições do tipo estão em tramitação (Brasil, 2024g). Identificou-se que desse total, em trinta casos o termo inteligência artificial está em uso apenas na justificação, principalmente quando se argumenta que a sociedade atual dispõe da inteligência artificial e outras tecnologias, sem que a proposição tenha qualquer relação com o termo.

[5] A busca foi realizada no endereço: https://www6g.senado.leg.br/busca/?portal=Atividade+Legislativa&q=intelig%C3%AAncia+artificial.
[6] A busca foi realizada no endereço: https://www.camara.leg.br/buscaProposicoesWeb/pesquisaSimplificada.

De janeiro a junho de 2024 foram apresentados na Câmara dos Deputados vinte e sete Projetos de Lei que estão diretamente relacionados ao desenvolvimento e/ou uso de inteligência artificial. As propostas são variadas, e vão desde a busca pelo estabelecimento de normas gerais para a área (a exemplo do PL 1.797/24) a indicações de mudanças no Código Penal (PL 477/24), Código Civil (PL 390/2024), Marco Civil da Internet (PL 842/2024) e outras leis especiais, para o acréscimo de disposições relativas à IA (Brasil, 2024g).

Com a ascensão da IA generativa, ampliou-se as possibilidades de manipulação e geração de imagens, sons e vídeos. Isso tem repercutido na apresentação de proposições que buscam coibir práticas danosas de uso nesse contexto, a exemplo do PL n. 2.506/24, PL n. 1.758/24 e PL n. 5928/23, que expressamente fazem referência, na justificação, ao uso da IA generativa na criação de conteúdos (Brasil, 2024f).

Diante do exposto, sobressai que a busca por estabelecer limites e parâmetros ao desenvolvimento e uso da inteligência artificial, na realidade brasileira, teve por consequência a apresentação de dezenas de Projetos de Lei. Algumas proposições são gerais e traçam diretrizes amplas, enquanto outras se propõem a alterar pontualmente a legislação já existente, em áreas específicas. Apesar das divergências entre as propostas, nota-se que existem repetições de conteúdos que já foram apresentados; assim, o esforço utilizado na elaboração de um novo Projeto poderia ser direcionado para aprimorar o debate daqueles que já estão em tramitação, uma vez que as proposições não são estanques e podem passar por emendas. Fato é que a aprovação de uma legislação robusta sobre o tema é crucial para o desenvolvimento da área com segurança e responsabilidade.

Em março de 2024, o Parlamento Europeu aprovou o chamado “AI Act”, que objetiva promover inteligência artificial confiável e centrada no ser humano, ao mesmo tempo em que garante um alto nível de proteção da saúde, da segurança, dos direitos fundamentais e também o apoio à inovação (European Parliament, 2024). A maior parte do “AI Act” será plenamente aplicável vinte e quatro meses após a sua entrada em vigor, mas existem ressalvas que permitirão uma melhor adaptação das partes atingidas, a exemplo das obrigações para sistemas de alto risco, em que o prazo será estendido para trinte e seis meses (European Parliament, 2024b). Em razão do nível de abrangência, essa regulação é um marco que servirá de referência para países que buscam estabelecer normas gerais sobre a IA, a exemplo do Brasil.

3 ÉTICA E TRANSPARÊNCIA NOS SISTEMAS DE IA

Segundo Marcondes (2007, p. 8), a ética, em um sentido amplo, diz respeito à determinação do que é certo ou errado, bom ou mau, permitido ou proibido, de acordo com um conjunto de normas ou valores adotados historicamente por uma sociedade. A ética, portanto, não pode ser vista dissociada de uma realidade sociocultural. Para o autor, "a ética aborda, centralmente, nossa vida concreta, nossa prática cotidiana" e não há uma resposta única ou geral para todas as questões éticas (Marcondes, 2007, p. 9-10; 15).

Ao refletir sobre a ética no contexto da inteligência artificial é preciso considerar especificidades da área. De acordo com Waelen, o campo emergente da ética da IA é diferente de outros campos da ética aplicada. No centro da sua atenção estão as formas pelas quais os humanos são afetados por essa tecnologia, que trouxe possibilidades de ação radicalmente novas (Waelen, 2022, p. 13).

Zhou e Chen (2022, p. 3) indicam que as teorias éticas centradas na ação concentram-se no que um agente deve fazer e em como determinar a ação correta em circunstâncias específicas, enquanto as teorias éticas centradas no agente estão focadas no ser e no bom caráter moral. Para os autores, a ética da IA é centrada tanto na ação quanto no agente. Assim, existe a preocupação com o comportamento dos humanos à medida que projetam, constroem e utilizam modelos artificialmente inteligentes, bem como há a preocupação a respeito do comportamento dos agentes de IA.

Waelen (2022, p. 2-4) defende que a ética da IA deve ser vista a partir das lentes de uma teoria crítica, pois isso possibilita um entendimento interdisciplinar das questões éticas, oferece um alvo para a mudança, ajuda a identificar implicações sociais e políticas de uma tecnologia e ajuda o campo a avançar. Essa compreensão, do ponto de vista da autora, se faz necessária em razão do surgimento de novas relações que podem afetar a autonomia do indivíduo e exacerbar as estruturas de poder existentes na sociedade.

O crescimento da inteligência artificial e da robótica nos últimos anos destacou a necessidade premente de se criar uma estrutura legal adequada. Os domínios mais importantes do tema vão se firmando lentamente, sendo o objetivo mais urgente a definição dos fundamentos éticos que sustentam a expansão da IA (Księżak; Wojtczak, 2023, p. 1).

Huang et al. (2023, p. 800-801) sugerem que a investigação do tema ainda está na sua infância. Os autores indicam que para resolver os problemas éticos da área, primeiro é preciso

reconhecer e compreender os potenciais problemas/riscos éticos que a IA pode trazer, para depois ser possível formular adequadamente princípios e regras éticas a serem seguidos.

Segundo Ryan e Stahl (2021, p. 62-63), não é controverso afirmar que o interesse pela ética e a IA é um fenômeno global. A atenção dada ao tema pela literatura cresceu a tal ponto de ser difícil mantê-lo atualizado. Para os autores, as características da IA que dão origens às preocupações éticas relacionam-se com a capacidade desta de agir e aprender de forma mais ou menos autônoma, a depender do "*input*" fornecido. Ademais, os sucessos e as conquistas recentes de algumas aplicações de IA fez com que os estudos ganhassem mais força.

De acordo com Hartmann Peixoto, a IA está situada em um espectro de relevância que inclui governos, indústria e academia. Assim, é indispensável e urgente o conhecimento de suas potencialidades, riscos e desafios, de modo que o desenvolvimento e aplicação desses sistemas devem atender a diretrizes éticas que sejam trabalhadas, refletidas e debatidas (Hartmann Peixoto, 2020c, p. 20).

Para Littman et al. (2021, p. 45), a linha divisória entre academia e indústria se torna cada vez mais tênue no campo da IA, com isso questões sociais e éticas vêm à tona. Organizações como a *Partnership on AI (PAI)*, que tem como foco a elaboração de boas práticas no desenvolvimento da IA, recebe financiamento de indústrias e universidades empenhadas nesse objetivo comum. Por outro lado, há preocupações acerca da atuação de empresas que monitoram e governam seu próprio comportamento ético, já que poderiam facilmente retirar o apoio de qualquer grupo ou iniciativa cujas conclusões entrem em conflito com os seus interesses comerciais a curto prazo (Littman et al., 2021, p. 46-47).

3.1 Desafios e questões éticas relacionadas à IA

À medida que a IA se torna mais sofisticada e adquire a capacidade de realizar tarefas humanas mais complexas, cresce a dificuldade em monitorar, validar, prever e explicar o seu comportamento. Como consequência, as preocupações éticas e os debates sobre os princípios e valores que devem orientar o desenvolvimento e a implantação da IA se amplificam (Zhou; Chen, 2022, p. 2).

É possível categorizar as questões éticas da IA a partir de diferentes perspectivas. Huang et al. (2023, p. 806) propõem uma classificação das questões éticas nos níveis individual, social e ambiental.

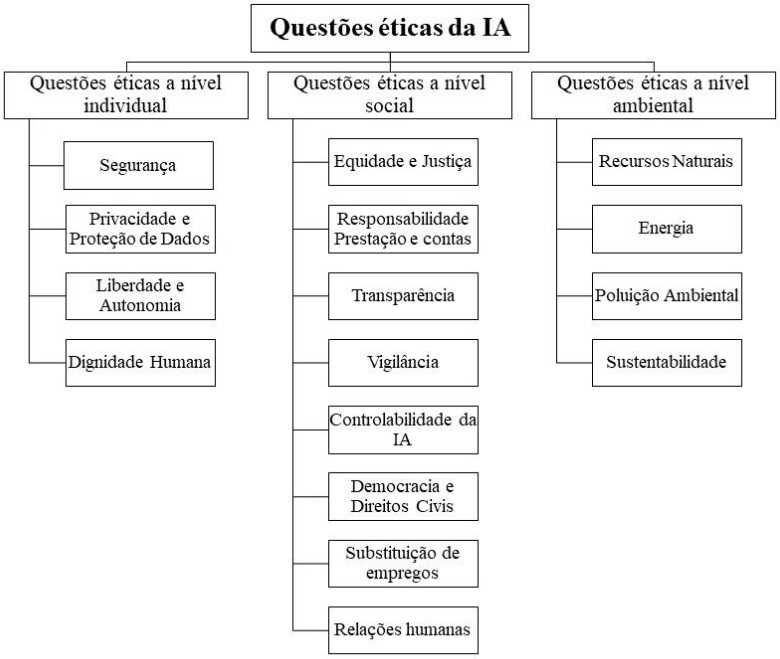
O nível individual das questões éticas considera as consequências indesejáveis para os seres humanos individualmente, os seus direitos e o seu bem-estar. Esse recorte reputa que aplicação da IA pode apresentar riscos para a segurança dos indivíduos, a exemplo de acidentes pessoais envolvendo carros autônomos. A privacidade é outro dilema apontado, já que os sistemas requerem uma enorme quantidade de dados para um bom funcionamento e existem sérios riscos associados ao recolhimento e tratamento de dados privados. Ademais, quando a tomada de decisões baseada em IA é amplamente adotada na vida diária, há o perigo de restrição da autonomia individual (Huang et al., 2023, p. 806).

Quanto ao nível social, o foco das questões éticas dá-se quanto aos reflexos que a IA traz para a sociedade e para o bem-estar das comunidades e nações. Sob essa categorização, discute-se equidade e justiça, responsabilidade e prestação de contas, transparência, vigilância, controlabilidade da IA, democracia, substituição de empregos e relações humanas (Huang et al., 2023, p. 806).

A nível ambiental, as questões éticas centram-se nos impactos da IA para o meio ambiente e o planeta. Assim, apesar da IA trazer comodidade às tarefas diárias, há um custo para o planeta, pois o uso desses sistemas requer a implantação de um grande número de terminais de hardware, incluindo chips, sensores, mecanismos de armazenamento, etc. A produção desses dispositivos consome recursos naturais, os quais, ao final de um ciclo de vida, poderão gerar poluição ambiental aos serem descartados. Além disso, os sistemas de IA geralmente requerem um poder computacional considerável, o que acarreta um alto consumo de energia (Huang et al., 2023, p. 806).

A categorização das questões éticas propostas por Huang et al. (2023) foi sintetizada conforme o esquema abaixo reproduzido:

Figura 12 - Questões éticas da IA em diferentes níveis



Fonte: Huang et al.,2023. Tradução própria.

Outras classificações das questões éticas da IA incluem: tópicos éticos relacionados às características da IA, os riscos éticos causados por fatores humanos, questões baseadas na implantação da IA e seus impactos na sociedade, no Direito, no sistema financeiro, além de outras classificações criadas pela literatura especializada (Huang et al., 2023, p. 802).

Segundo Alice Xiang, os modelos de IA são como espelhos que refletem padrões da sociedade, sejam eles justos ou injustos; e também as visões de mundo de seus criadores, que podem ser enviesadas (Xiang, 2024, p. 250). A problemática dos vieses é apontada com recorrência pela literatura como uma questão ética que deve ser enfrentada e debatida no desenvolvimento e uso de sistemas de IA.

Para Kahneman, Sibony e Sunstein (2022, p. 161), a palavra “viés” possui um significado amplo. O viés pode se referir a um mecanismo psicológico e ao erro que esse mecanismo tipicamente produz. Além disso, a palavra é frequentemente usada para sugerir que alguém é tendencioso contra determinado grupo ou que tende a preferir uma conclusão particular.

Há inúmeros casos registrados de algoritmos que perpetuaram disparidades de raça ou de gênero. A visibilidade desses casos explica a crescente preocupação com o viés na tomada

de decisão algorítmica. Tais disparidades podem vir à tona se determinado modelo preditor utiliza variáveis correlacionadas à raça ou se a fonte dos dados com os quais o algoritmo é treinado for enviesada (Kahneman; Sibony; Sunstein, 2022, p. 131-132).

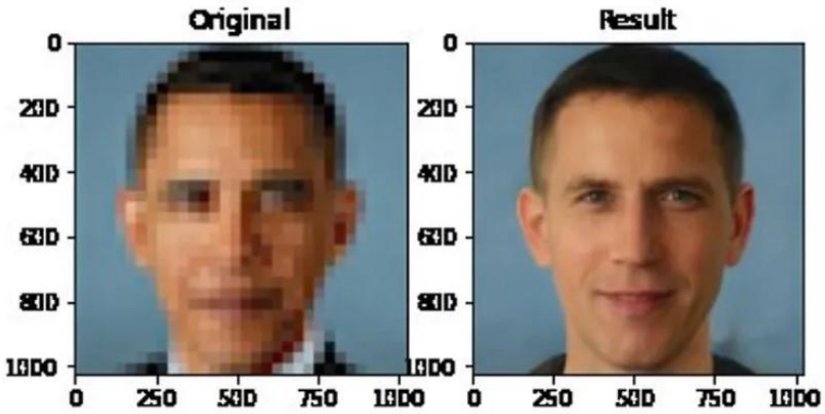
Em 2016, um estudo independente divulgado pela Pro Publica indicou que o sistema de apoio à decisão de algumas cortes dos EUA, chamado de Correctional Offender Management Profiling for Alternative Sanctions – COMPAS, era racialmente tendencioso. Ao realizar a classificação de risco de reincidência, o algoritmo tinha uma probabilidade particular de sinalizar falsamente os réus negros como futuros criminosos, e os réus brancos eram sistematicamente rotulados como de baixo risco. Muitos tribunais americanos adotaram o sistema antes mesmo de testar seu real funcionamento (Angwin et al., 2016).

A empresa desenvolvedora do software de avaliação de risco, a Northpointe, utilizava no modelo um conjunto de pontuações derivadas de 137 perguntas, que eram respondidas pelos réus ou extraídas de registros criminais. A raça não era uma das questões, contudo, a pontuação incluía itens que poderiam ser correlacionados com a cor da pele. Os cálculos que transformavam os dados subjacentes numa pontuação não foram revelados, sob o argumento de sigilo comercial (Angwin et al., 2016).

Além da possibilidade de que os dados escolhidos para um modelo reflitam preconceitos existentes ou estejam correlacionados a discriminações preexistentes, é possível que os dados coletados não sejam representativos da realidade. Caso um algoritmo de aprendizagem profunda receba mais fotos de rostos de pele clara do que de rostos de pele escura, por exemplo, o sistema de reconhecimento facial resultante seria inevitavelmente pior no reconhecimento de rostos de pele mais escura (Hao, 2019).

A imagem abaixo ilustra a reprodução de vieses por um sistema de IA. Segundo James Vicent (2020), o programa PULSE, que utiliza o algoritmo StyleGAN para gerar rostos com alta resolução a partir de imagens pixeladas, foi alvo de discussões após a inserção de uma imagem de baixa resolução de Barack Obama, o primeiro presidente negro dos Estado Unidos, gerar como saída a foto de um homem branco.

Figura 13 - Imagem gerada através do algoritmo StyleGAN a partir de foto pixelada do ex-presidente Barack Obama



Fonte: Vicent, 2020.

No PULSE, a transformação de uma imagem de baixa resolução para uma de alta resolução ocorre por meio do preenchimento de lacunas via aprendizado de máquina. A geração de rostos brancos de forma tendenciosa ocorre devido aos dados utilizados para o treinamento do algoritmo serem frequentemente distorcidos em direção a um único grupo demográfico, porém, quando o programa lida com dados que não estão nesse grupo demográfico, ele tem um desempenho ruim (Vicent, 2020).

Segundo Cozman e Kaufman (2022, p. 201), é recente a sensibilidade de pesquisadores e da sociedade para o problema dos vieses nos dados. Os autores indicam que durante anos diversas bases de dados tendenciosas foram utilizadas para desenvolver e treinar algoritmos de IA. O viés algorítmico é difícil de ser detectado quando atrelados a sistemas que estão protegidos por sigilo comercial e também quando a composição do modelo é mais sofisticada, envolvendo treinamentos em múltiplas bases de dados (Cozman; Kaufman, 2022, p. 202).

Sobre os vieses algorítmicos, a Estratégia Brasileira de Inteligência Artificial parte do pressuposto de que a IA não deve criar ou reforçar preconceitos, principalmente os relacionados a características sensíveis como raça, etnia, gênero, nacionalidade, renda, orientação sexual, deficiência, crença religiosa e inclinação política. Nesse sentido, as organizações que desenvolvem e usam sistemas de IA devem estar cientes dos princípios balizadores de seus sistemas e verificar periodicamente se estes estão sendo respeitados (EBIA, 2021, p. 22).

Demandas éticas na relação entre o Direito e a inteligência artificial, pela característica multidisciplinar, estão relacionadas ao surgimento de situações limite, seja com a execução de

atividades fruto de sistemas de aprendizagem de máquina, interconexão do raciocínio jurídico com o raciocínio matemático, indução de preferências e outros. No mecanismo de apoio à decisão, por exemplo, decisões tendenciosas, que aprofundem preconceitos, são incompatíveis com diretrizes de pesquisa, desenvolvimento e uso em um ambiente democrático e de conscientização de direitos fundamentais (Hartmann Peixoto, 2020c, p. 30).

O Painel de Projetos de IA no Poder Judiciário - 2023, elaborado no âmbito do Programa Justiça 4.0, divulgou, a partir das respostas fornecidas por 91 tribunais e 3 conselhos de justiça (Conselho Nacional de Justiça, Conselho da Justiça Federal e Conselho Superior da Justiça do Trabalho), as preocupações éticas relacionadas ao uso de IA (CNJ, 2023). Na tabela reproduzida a seguir estão identificadas as principais preocupações:

Tabela 2 - Principais preocupações éticas relacionadas ao uso IA nos tribunais e conselhos

Preocupações éticas relacionadas ao uso de IA	**Quantidade de respostas**
Discriminação e viés nos resultados obtidos pelos modelos de IA por conta da base de treinamento do modelo.	90
Responsabilidade e *accontability* em caso de decisões equivocadas da IA.	87
Falta de transparência nas decisões tomadas pelos algoritmos de IA.	71
Falta de transparência e auditabilidade no processo de treinamento do(s) modelo(s).	65
Violação da privacidade das partes envolvidas nos processos judiciais.	59
Potencial substituição de profissionais humanos por sistemas automatizados.	30
Não sei	27

Tabela elaborada com base nos dados do Painel de Projetos de IA no Poder Judiciário - 2023 (CNJ, 2023).

Os problemas no subcampo da ética da IA são um microcosmo dos problemas sociais mais amplos, argumenta Alice Xiang (2024, p. 262). Do ponto de vista da autora, em razão da IA ser uma criação humana, é estabelecido um requisito moral mais forte para evitar o emprego

de uma IA que perpetue e consolide os problemas sociais existentes; e, de certa forma, tem-se mais controle sobre os sistemas do que sobre a sociedade em geral. Contudo, desenvolver modelos mais justos é uma tarefa difícil porque a IA reflete a sociedade e todas as suas complexidades (Xiang, 2024, p. 262).

3.2 Diretrizes e princípios éticos que orientam o desenvolvimento e o uso da IA

O relatório "*The One Hundred Year Study on Artificial Intelligence*" (AI100), divulgado pela primeira vez em 2016, e que teve sua última edição publicada em 2021, afirma que à medida que a IA cresceu em sofisticação e a sua utilização se tornou mais generalizada, os governos têm prestado cada vez mais atenção ao seu desenvolvimento e implantação. Mais de 60 países envolveram-se em iniciativas nacionais para normatizar o tema, mas poucos países avançaram definitivamente na regulação específica da IA, fora das regras diretamente relacionadas com utilização de dados (Littman et al., 2021, p. 37).

Vários grupos internacionais desenvolveram esforços e sugestões destinadas a gerar quadros políticos para o desenvolvimento e uso responsável da IA, resultando em iniciativas como os Princípios para a IA da OCDE - Organização para a Cooperação e Desenvolvimento Econômico (Littman et al., 2021, p. 38-39).

Os princípios estabelecidos em 2019 pela OCDE foram atualizados em 2024, são eles: crescimento inclusivo, desenvolvimento sustentável e bem-estar; direitos humanos e valores democráticos, incluindo justiça e privacidade; transparência e explicabilidade; robustez, segurança e proteção; e responsabilidade. Além dos princípios, a organização indica cinco recomendações para os gestores públicos: investir em Pesquisa e Desenvolvimento de IA; promover um ecossistema inclusivo que possibilite a IA; moldar um ambiente de governança e política interoperável e propício para IA; desenvolver capacidades humanas e preparar a transição para o mercado de trabalho; e cooperação internacional para uma IA confiável. Apesar de não ser membro da OCDE, o Brasil se comprometeu com a adoção dos princípios e das recomendações da organização (OCDE, 2024).

A União Europeia tem estado mais ativa em iniciativas concretas, incluindo o Regulamento Geral de Proteção de dados - RGPD e o Quadro de Aspectos Éticos da IA, Robótica e Tecnologias Relacionadas, que propõe a criação de órgãos nacionais de supervisão e a designação de tecnologias de alto risco (Littman et al., 2021, p. 39). Ademais, em 2024 o Parlamento Europeu aprovou o chamado "AI Act", regulação já mencionada no segundo capítulo do trabalho e que possui relação com os princípios éticos sugeridos no âmbito do bloco.

As diretrizes éticas da IA podem se destinar a uma série de partes interessadas, como gestores de políticas públicas, usuários, desenvolvedores, educadores, organizações da sociedade civil, entidades de classe e outros. Como consequência, muitas das diretrizes são demasiadamente amplas em termos de cobertura, ou seja, fornecem indicações para muitos grupos diferentes, o que torna mais difícil a compreensão das orientações (Ryan; Stahl, 2021, p. 62-64). Considerando tal aspecto, a presente seção tem como objetivo analisar diretrizes éticas voltadas aos usuários e desenvolvedores de sistemas de IA, com foco principalmente na utilização desses sistemas pelo Direito.

Segundo Rosalie Waelen (2022, p. 1-2), a abordagem para a análise da ética de sistemas de IA com base em princípios é uma das mais dominantes. Nos últimos anos, inúmeras iniciativas desenvolveram conjuntos de princípios e diretrizes para garantir um desenvolvimento e uso desejável da IA, que funcionam como uma espécie de *soft law,* em razão de não serem juridicamente vinculativos e adotarem caráter orientador.

Mesmo sem as características do chamado *hard law*, isto é, normas vinculativas que definem condutas permitidas ou proibidas, as diretrizes éticas possuem natureza persuasiva e têm influência prática na tomada de decisões em determinados domínios (Jobin; Ienca; Vayena, 2019, p. 389).

Jobin, Ienca e Vayena (2019, p. 389) afirmam que há uma convergência global emergente em torno de cinco princípios éticos: transparência, justiça e equidade, não maleficência, responsabilidade e privacidade; com divergências em relação à forma como esses princípios são interpretados.

Em um estudo posterior, Waelen (2022, p. 6) reforça a ideia de convergência entre os princípios propostos por diferentes partes - como acadêmicos, empresas e decisores políticos - em que é possível a identificação de *clusters* que contém os princípios da transparência, justiça e equidade, não maleficência, beneficência, responsabilidade, *accountability*, privacidade, liberdade e autonomia, confiança e solidariedade (Waelen, 2022, p. 6).

Segundo Waelen (2022, p. 7), o princípio mais proposto para uma IA ética é o da transparência. A ideia por trás deste princípio é a de que uma IA transparente minimizaria os potenciais danos causados pelos sistemas que utilizam essa tecnologia, melhoria a interação humano-IA e aumentaria a confiança nos modelos. Do ponto de vista da autora, o que torna a transparência tão valiosa é que ela concede ao indivíduo a capacidade de saber o que está acontecendo com os seus dados e como a IA os afeta.

A transparência pode ser entendida de duas maneiras: a transparência da própria tecnologia de IA e a transparência das organizações de IA que a desenvolvem e utilizam-na. Os desenvolvedores de sistemas de IA precisam garantir a transparência, porque através dela protege-se direitos humanos fundamentais, a privacidade, a dignidade, a autonomia e o bem-estar humano. As organizações devem ser transparentes acerca dos benefícios, danos e os resultados potenciais, a fim de possibilitar que os consumidores façam escolhas informadas sobre a partilha dos seus dados (Ryan; Stahl, 2021, p. 66).

De acordo com Hartmann Peixoto, deve haver transparência e divulgação responsável para que as pessoas entendam os resultados baseados em sistemas de IA e possam questioná-los. O autor recomenda que um modelo de transparência e auditabilidade deve, sempre que possível, envolver um processo de certificação de boas práticas a ser ofertado pelos entes envolvidos - academia, indústria e governo (Hartmann Peixoto, 2020c, p. 145).

Para Arrieta et al. (2020, p. 85), a transparência se refere à característica de um modelo ser, por si só, compreensível para um ser humano. Os autores dividem os modelos transparentes em três categorias: simuláveis, decomponíveis e algoritmicamente transparentes. A simulabilidade denota a capacidade de um modelo ser simulado ou pensado estritamente por um ser humano, assim, sistemas baseados em regras simples, mas com um grande número delas, ficariam de fora dessa característica. A decomponibilidade significa a capacidade de explicar cada uma das partes de um modelo (entrada parâmetro e cálculo). Por fim, a transparência algorítmica é vista como a capacidade do usuário de compreender o processo seguido pelo modelo para produzir qualquer saída a partir dos dados de entrada (Arrieta et al., 2020, p. 87-88).

Segundo Lage e Hartmann Peixoto, a transparência, em uma perspectiva regulatória, deve ter requisitos dependentes do contexto e baseados nos riscos à segurança, justiça e privacidade. Em síntese, deve-se exigir mais transparência para decisões baseadas em IA que tenham um grande impacto em alguém além do próprio decisor (Lage, Hartmann Peixoto, 2021, p. 285).

A Estratégia Brasileira de Inteligência Artificial - EBIA, quanto ao princípio da transparência, frisa a necessidade de adoção de medidas para garantir a compreensão dos processos associados à tomada de decisões automatizadas, tornando possível identificar vieses envolvidos no processo decisório e desafiar as referidas decisões, quando cabível. Segundo o documento em questão, são elementos-chave da discussão internacional sobre o tema: i) a ideia

de que os sistemas devem ser centrados no ser humano - *human-centric AI*; e ii) a necessidade de que tais sistemas sejam confiáveis - *trustworthy AI* (EBIA, 2021, p. 18).

No subcampo da denominada *Explainable Artificial Intelligence (*XAI), Arrieta et al. (2020, p. 84) esclarecem que há o uso intercambiável de interpretabilidade e explicabilidade na literatura, o que dificulta o estabelecimento de bases comuns. Para os autores, a interpretabilidade refere-se a uma característica passiva de um modelo, referindo-se ao nível em que o sistema faz sentido para um ser humano observador, de modo que essa característica também é expressa como transparência. A explicabilidade, por sua vez, pode ser vista como um atributo ativo do modelo, denotando qualquer ação ou procedimento realizados com a intenção de esclarecer ou detalhar as funções internas de um modelo.

Ainda quanto à explicabilidade, Ryan e Stahl (2021, p. 66) indicam que as organizações que desenvolvem aplicações de IA devem ser capazes de explicar de forma inteligível os dados que entram, os dados que saem, o que os seus algoritmos fazem e o objetivo a ser alcançado ao fazê-lo. Caso exista uma tensão entre desempenho e explicabilidade, esta deve ser claramente identificada.

Os algoritmos de IA sofrem, em algum grau, com o problema da opacidade, em que por vezes é difícil obter informações sobre seu mecanismo interno de trabalho. Para enfrentar esta questão, o campo da inteligência artificial explicável (XAI) busca criar um conjunto de técnicas que produzam modelos mais explicáveis, mantendo altos níveis de desempenho. Para além do desafio de desmitificar as caixas pretas (termo utilizado quando o sistema não revela nada sobre o seu design interno, estrutura e implementação), a XAI é essencial para que os usuários entendam, confiem adequadamente e gerenciem com eficácia os resultados da IA (Adadi; Barrada, 2018, p. 1-5).

Entre as razões para a necessidade de explicar os sistemas de IA estão a justificação de resultados específicos geradas pelo modelo; o controle sobre vulnerabilidades e falhas dos sistemas, permitindo a correção de erros; a melhoria contínua da ferramenta, já que a melhor compreensão facilita o aprimoramento; e a geração de conhecimento, pois com a explicabilidade é possível aprender novos fatos e recolher informações sobre o desempenho da máquina (Adadi; Barrada, 2018, p. 5-6).

Segundo Bonat e Hartmann Peixoto (2020, p. 51), a explicabilidade está apoiada nas características de *accountability* e *auditability*, ou seja, a possibilidade de rastreamento do trajeto para a tomada de decisão objetivando sua prestação de contas e a fiscalização, com a

possibilidade de verificação e revisão dos processos, testes e ajustes para prevenir falhas futuras. Na visão dos autores, o cumprimento da explicabilidade não significa necessariamente que o sistema deva ser entendido nas suas minúcias por todos.

Para Morley et al. (2020, p. 2155), o princípio da explicabilidade se tornou vital na comunidade ética do *machine learning* porque, até certo ponto, está ligado aos princípios da autonomia, justiça, beneficência e não-maleficência. Para os autores, um sistema explicável é inerentemente mais transparente e, portanto, mais responsável em termos das suas propriedades de tomada de decisão.

Os usuários finais devem ser informados de que estão interagindo com um sistema de IA e a equipe desenvolvedora deve ser capaz de explicar e justificar o uso da sua IA. Os sistemas devem passar por auditorias internas e externas para garantir que é adequado à finalidade proposta, e as organizações devem permitir a análise e revisão independentes dos seus sistemas (Ryan; Stahl, 2021, p. 67).

O aspecto de que os indivíduos devem ter ciência de suas interações com sistemas de IA é apresentado pela EBIA, já que a oferta de informação quanto à existência de processos de tomada de decisões baseadas em IA é fundamental para o exercício do direito de revisão de decisões automatizadas previsto na LGPD, que inclui as decisões destinadas a definir o perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade (EBIA, 2021, p. 21).

Quanto ao princípio ético da dignidade, Ruster, Oliva-Altamirano e Daniell (2022, p. 2-3) propõem que ele seja visto a partir da chamada "lente da dignidade", cuja utilização se dá em contextos onde designers, desenvolvedores e outros profissionais estejam interessados em incorporar a dignidade prospectivamente e/ou verificar até que ponto a dignidade está inserida em um sistema. A orientação sobre dignidade, asseveram, está relacionada em respeitar o valor intrínseco dos seres humanos, como por exemplo, em ser claro quando um usuário está interagindo com uma IA e não com outro ser humano.

Os princípios da justiça e equidade estão relacionados com a igualdade de acesso à IA e os efeitos que o uso desses sistemas podem trazer. Existe a preocupação de que surja uma divisão digital entre os países que podem desenvolver e utilizar a IA e as partes do mundo que não têm acesso à tecnologia mais recente. O princípio da não discriminação tornou-se premente diante da constatação de que os softwares de IA podem ser tornar propagadores de preconceitos, de modo que o subcampo dos vieses algorítmicos tem recebido muito atenção. Assim,

algoritmos tendenciosos podem levar a resultados discriminatórios, e, portanto, violar a pretensão de uma IA justa e equitativa (Waelen, 2022, p. 8).

Para Hartmann Peixoto, o princípio da justiça substancial traz a ideia de que os sistemas de IA têm a responsabilidade ativa pela realização de justiça e o compromisso com a inclusão e a equidade. Em uma situação de envolvimento de sistema autônomo na tomada de decisões judiciais, por exemplo, deve ser fornecido uma explicação satisfatória auditável por uma autoridade humana competente, a fim de verificar se está sendo observado o compromisso de justiça (Hartmann Peixoto, 2020c, p. 143).

O princípio da solidariedade, apontado com menos frequência nas diretrizes existentes, diz respeito à distribuição justa dos benefícios e danos da IA. Isso implicaria que os benefícios da IA deveriam ser redistribuídos daqueles que são desproporcionalmente beneficiados, para aqueles que se revelam mais vulneráveis a ela - a exemplo dos desempregados devido à automação (Waelen, 2022, p. 8).

Em sentido semelhante, Jobin; Ienca; Vayena (2019, p. 396) expressam que o princípio da solidariedade remete à necessidade de uma rede de segurança social, com foco na redistribuição dos benefícios da IA, a fim de não ameaçar a coesão social e de respeitar pessoas e grupos potencialmente vulneráveis. A solidariedade é igualmente referenciada em relação às implicações da IA para o mercado de trabalho.

O princípio da solidariedade também está relacionado à perspectiva de não se prejudicar a manutenção de relacionamentos humanos morais e emocionais, de modo que os sistemas de IA devem atuar na promoção dos relacionamentos humanos e reduzir a vulnerabilidade e isolamento das pessoas (Hartmann Peixoto, 2020c, p. 148).

O debate acerca dos caminhos segundo os quais a IA deve seguir expressa a preocupação com a posição do ser humano frente ao (potencial) poder da IA. Portanto, o objetivo da ética nesse contexto é garantir que as tecnologias emergentes que prometem mudar a vida tal como a conhecemos o façam para melhor (Waelen, 2022, p. 10-11).

Na ética da IA, evitar danos aos seres humanos tem sido uma das maiores preocupações. O princípio da não-maleficência enfatiza que as organizações que desenvolvem e utilizam a IA devem incorporar o aconselhamento de autoridades legais e conselhos de ética em pesquisa para garantir que os dados sejam recuperados, analisados e empregados de uma maneira que não prejudique os indivíduos. Busca-se que a IA seja robusta, segura e protegida ao longo do

seu ciclo de vida, de forma a funcionar adequadamente e não representar riscos de segurança excessivos (Ryan; Stahl, 2021, p. 70).

Existem inúmeros debates sobre a responsabilidade da IA. Jobin, Ienca e Vayena (2019, p. 395) identificaram que as diretrizes éticas nomeiam como responsáveis pelas ações e decisões da IA atores muito diferentes. Entre os nomeados estão: desenvolvedores de IA, designers, instituições e indústria. Os autores ainda apontam que as divergências entre os documentos aumentam quanto à discussão se a IA deveria ser responsabilizada de forma semelhante à humana ou se os humanos deveriam ser sempre os únicos intervenientes responsáveis.

Com relação à prestação de contas (*accountability*), enfatiza-se a auditabilidade por meio da avaliação de algoritmos, dados e processos de design, de maneira a preservar a propriedade intelectual do modelo. As avaliações podem ser realizadas por auditores internos e externos; contudo, quando o sistema de IA afeta direitos fundamentais, deve sempre ser auditado por um terceiro externo. Também está compreendida na ideia de *accountability* a atividade de reportar ações ou decisões que produzem determinado resultado por parte do sistema, atentando-se para a identificação, avaliação, documentação e minimização dos seus impactos negativos (Arrieta et al., 2020, p. 105).

Considerando a importância dos dados para os modelos de inteligência artificial, os princípios da representação substancial no desenvolvimento, da autenticidade de *datasets* e da privacidade dos dados, evidenciados por Hartmann Peixoto, se destacam. O princípio da representação substancial no desenvolvimento é verificado *a priori* e busca uma paridade de representação no *dataset*, na fase de treinamento, e a proteção contra preconceitos. A autenticidade de *datasets* indica que deve existir uma autenticidade na própria construção do conjunto de dados. O princípio da privacidade dos dados, na visão do autor, envolve a salvaguarda de dados privados, dados sensíveis e o estabelecimento de um sistema de governança de dados (Hartmann Peixoto, 2020c, p. 141-147).

A privacidade na ética da IA pode ser vista tanto como um valor a ser defendido como um direito a ser protegido. Esse princípio é frequentemente associado à proteção de dados e à segurança de dados. Algumas fontes o associam à liberdade ou à confiança (Jobin; Ienca; Vayena, 2019, p. 395). A importância da privacidade se faz presente durante todo o ciclo de vida de um modelo e, por não ser um princípio exclusivo dos sistemas de IA, uma vez que está presente em muitos outros produtos de software, as práticas de privacidade podem ser herdadas de processos que já existem em uma organização (Arrieta et al., 2020, p. 103).

Os princípios da liberdade e autonomia relacionam-se, em algumas diretrizes, à liberdade de expressão e à autodeterminação informacional. Alguns documentos referem-se à autonomia como uma liberdade positiva, a exemplo da liberdade de retirar o consentimento. Outros documentos centram-se na liberdade negativa, a exemplo da abstenção de experimentação tecnológica (Jobin; Ienca; Vayena, 2019, p. 395).

Acredita-se que a liberdade e a autonomia sejam promovidas através da transparência e da IA previsível, aumentando ativamente o conhecimento dos indivíduos sobre a IA e deixando de coletar e divulgar dados na ausência de consentimento informado (Jobin; Ienca; Vayena, 2019, p. 395).

A confiança é uma característica desejável na relação entre a tecnologia e aqueles que a utilizam ou estão sujeitos a ela. Ao confiar num sistema de IA, espera-se que o poder que a tecnologia pode exercer sobre um indivíduo não seja mal utilizado. Uma relação de poder confiável é aquela em que A detém poder sobre B, sem que B precise se preocupar com a possibilidade de A tirar vantagem dessa situação. O clamor por uma IA confiável busca garantir que esta não possa exercer poder de forma arbitrária, prejudicial ou excessiva (Waelen, 2022, p. 10).

Uma cultura de confiança nas potenciais utilizações da IA é indispensável para que esta cumpra seus objetivos. Além disso, um ambiente de confiança entre cientistas e engenheiros contribui para a realização dos objetivos organizacionais (Jobin; Ienca; Vayena, 2019, p. 395).

Jobin, Ienca e Vayena (2019, p. 396) alertam que divergências conceituais revelam incertezas quanto a quais princípios éticos devem ser priorizados e como os eventuais conflitos entre princípios devem ser resolvidos. A necessidade de conjuntos de dados cada vez maiores e mais diversificados para imparcializar a IA pode conflitar com o requisito de dar aos indivíduos um maior controle sobre os seus dados e sua utilização, explicam os autores.

Ao longo do ciclo de vida do sistema de IA diferentes ênfases nos princípios éticos são observadas. Na fase de aquisição de dados, por exemplo, a privacidade é um princípio fundamental, enquanto que na fase de construção de aplicações, as partes envolvidas estão mais interessadas na transparência do modelo. Portanto, a implementação dos princípios éticos pode ocorrer nas distintas fases do clico de vida do modelo e a educação sobre a ética da IA pode ser útil para uma melhor compreensão da área (Zhou; Chen, 2022, p. 11).

O conjunto de princípios forma a estrutura axiológica da IA ética e deve orientar a formação e a interpretação normativa para o desenvolvimento e uso da IA no Direito, indica Hartmann Peixoto. Para se falar em robustez, solidez, confiança e competitividade é preciso levar em conta a dimensão ética e a capacidade de impacto da IA no Direito. Os níveis de impactos influenciam a concretização dos princípios e também as justificativas de preponderância entre princípios (Hartmann Peixoto, 2020c, p. 161).

3.3 Direcionamentos para a implementação da ética no campo da IA

Uma avaliação cuidadosa de determinado sistema de IA é recomendada para assegurar que os algoritmos desconsiderem *inputs* inadmissíveis, além de ser aconselhada a realização de testes para verificar se discriminam com base na raça, no gênero, contra membros de grupos desfavorecidos ou apresentam outros resultados indesejáveis (Kahneman; Sibony; Sunstein, 2022, p. 325-326).

Kahneman, Sibony e Sunstein (2022, p. 326) possuem uma visão otimista acerca do uso de sistemas de IA. Do ponto de vista dos autores "embora um algoritmo preditivo em um mundo incerto dificilmente seja perfeito, ele pode ser muito menos imperfeito do que o julgamento humano ruidoso e frequentemente enviesado". O ruído, alertam, é a variabilidade indesejada em julgamentos que deveriam ser idênticos (Kahneman; Sibony; Sunstein, 2022, p. 351).

A IA pode resultar em progresso tecnológico, mas isso não garante progresso social. À vista disso, para melhorar a o campo da inteligência artificial é necessário um esforço interdisciplinar. Não apenas os especialistas em ética, mas também os desenvolvedores de *software*, cientistas de dados, decisores políticos e os juristas precisam estar envolvidos para concretizar o objetivo da IA ética (Waelen, 2022, p. 12-13).

Os instrumentos éticos da inteligência artificial, como princípios, estruturas, diretrizes, políticas e ferramentas, possibilitaram novos debates acerca de quais questões devem estar no centro do desenvolvimento dos sistemas. Embora tais instrumentos sirvam como um passo necessário para abordar as preocupações éticas, colocar essas orientações em prática é um desafio conhecido, refletido em apelos crescentes por mais trabalhos que encontrem formas de operacionalizar esse conhecimento (Ruster; Oliva-Altamirano; Daniell, 2022, p. 1).

Aplicar a ética ao desenvolvimento do *machine learning* é uma questão em aberto que pode ser resolvida de diversas maneiras, em diferentes escalas e em distintas situações, reconhecem Morley et al. (2020, p. 2152). Nem todos os princípios terão a mesma importância

em todos os contextos e essa ideia deve ser priorizada no desenvolvimento de ferramentas e métodos, de modo a oportunizar um espaço para tal flexibilidade (Morley et al., 2020, p. 2155).

Existem muitas barreiras entre as aspirações de se criar uma IA responsável e a tradução dessas aspirações em aspectos práticos concretos. A tarefa de estreitar as lacunas entre princípios e práticas é crítica para o futuro desta tecnologia, digna de atenção dos governos, das organizações desenvolvedoras e do público em geral. Schiff et al. (2021, p. 81-82) analisam seis possíveis explicações para a existência de tais lacunas, sintetizadas no quadro abaixo:

Figura 14 - Lacunas entre princípios e práticas

Fonte: Schiff et al.,2021. Tradução própria.

A primeira explicação, chamada de dilema dos incentivos, está relacionada ao fato de que os valores, motivações e incentivos que orientam as empresas podem não estar suficientemente alinhados com a utilização responsável da IA. Algumas organizações são movidas pela lógica econômica ou financeira, em que a busca por lucros está acima das preocupações legais e sociais. É possível ponderar valores concorrentes por meio da autorregulação coletiva da indústria, mudanças na educação e treinamento de engenheiros e

adoção de um foco mais amplo em benefícios para as partes interessadas (Schiff et al., 2021, p. 82-83).

Os impactos da IA no bem-estar são mais complexos do que às vezes se supõe. Os desenvolvedores muitas vezes se concentram em produtos únicos e nos danos físicos que estes podem causar, ao invés de tipos mais amplos de danos sociais, emocionais ou econômicos. Para além de um enfoque restrito, a IA deve ser entendida como algo que tem influência em múltiplos aspectos do bem-estar humano e social, tais como direitos humanos, desigualdades sociais, coesão social e política, saúde, e, em um nível mais abrangente, pode ter impacto nos ecossistemas naturais e na vida animal (Schiff et al., 2021, p. 83).

A divisão disciplinar resulta da pluralidade de conhecimentos e profissionais requeridos para a construção de uma IA responsável, que vai além dos papeis exercidos pelos engenheiros e cientistas computação. O problema das "muitas mãos" está associado à distribuição da responsabilidade pela gestão da IA. A separação funcional de especialistas técnicos e não técnicos em uma organização limita o potencial de comunicação eficaz e cria lacunas nos deveres pelas abordagens da qualidade e segurança de um produto (Schiff et al., 2021, p. 84 - 85).

Na perspectiva da governança do conhecimento, exige-se que o conhecimento coletado dentro da organização seja armazenado de forma a ser facilmente recuperável pelas equipes apropriadas, o que reforça a necessidade de se pensar de forma holística e com uma metodologia adequada. Quanto ao problema da abundância de ferramentas, tem-se que a proliferação excessiva de procedimentos dificulta a tarefa de classificação e avaliação da utilidade do instrumento, prejudicando a correção de falhas (Schiff et al., 2021, p. 85-86).

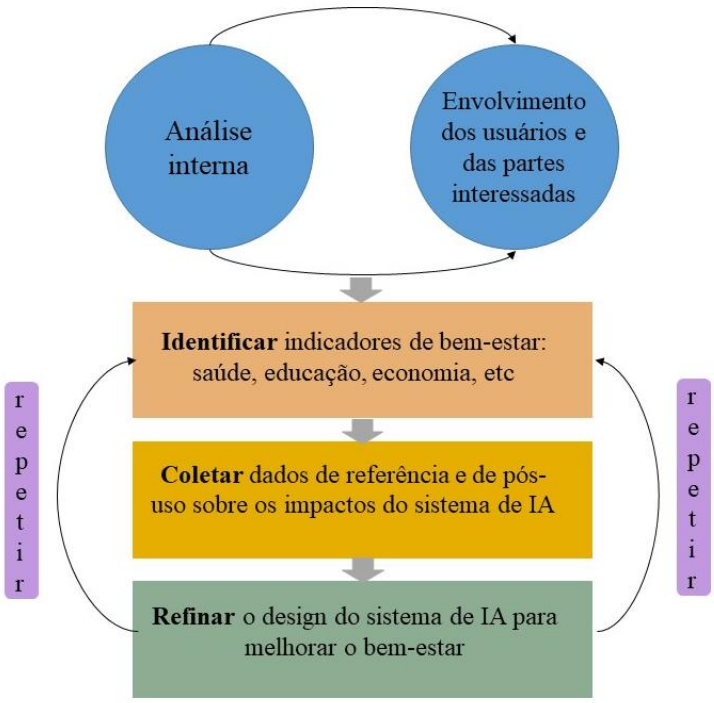
De acordo com Schiff et al. (2021, p. 86), as avaliações de impacto são uma estratégia promissora para alcançar critérios que possibilitam estreitar a lacuna entre princípios e práticas. As avaliações de impacto, que historicamente têm sido usadas em direitos humanos e em circunstâncias regulatórias, podem ser aplicadas à IA. A medição do impacto permite a monitorização dos riscos, a criação de um ambiente mais seguro para o investimento, a promoção da responsabilização e transparência e, em geral, aumenta as perspectivas de inovação pró-social.

Uma avaliação de impacto para a IA responsável, centrada no bem-estar do indivíduo, envolve: i) análise interna, ii) envolvimento dos usuários e das partes interessadas e iii) coleta de dados. Na primeira etapa é feito um dimensionamento dos danos, riscos e utilizações

intencionais e não intencionais de um sistema de IA. O contato dos desenvolvedores com os usuários ocorre na segunda etapa, em que se determina como o sistema afeta o bem-estar pessoal. A terceira etapa é caracterizada pela coleta de dados por meio de pesquisas com usuários, grupos, fontes de dados disponíveis publicamente ou diretamente do sistema (Schiff et al., 2021, p. 86-87).

Schiff et al. (2021, p. 87) alertam que aplicar uma avaliação de impacto pode parecer um exercício abstrato para aqueles que ainda não a fizeram, mas esta pode contribuir para a melhoria contínua dos sistemas. Abaixo foi reproduzido um esquema que demonstra o funcionamento de uma avaliação de impacto.

Figura 15 - Avaliação de impacto esquematizada



Fonte: Schiff et al., 2021. Tradução própria.

A Estratégia Brasileira de Inteligência Artificial adota o entendimento de que é necessário aprofundar o estudo dos impactos da IA em diferentes setores, a fim de evitar ações regulatórias que possam desnecessariamente limitar a inovação, de modo que uma avaliação de impacto contextual é útil para esse objetivo (EBIA, 2021, p. 22).

Ao elaborar uma arquitetura ética para o desenvolvimento e o uso de IA no Direito, Hartmann Peixoto (2020c, p. 153) propõe a implementação de estratégias e táticas para a obtenção de salvaguardas éticas na área. Entre as estratégias estão o esclarecimento sobre as formas de proteção contra preconceitos; a explicação sobre as formas de proteção aos grupos vulneráveis; a definição de medidas de segurança ativas e passivas, contingências e controle de erros específicos; e o estabelecimento de comunicação social para a apresentação de benefícios, melhores usos e resultados.

Entre as táticas elencadas por Hartmann Peixoto estão a orientação para que os sistemas de IA sejam facilitadores de uma sociedade democrática, equitativa, apoiando a ação do utilizador e a promoção dos direitos fundamentais; a consideração de que todo desenvolvimento deve ter metodologia clara de modo a evitar *bias* e a permissão de supervisão humana devidamente registrada. O autor esclarece que as recomendações estratégicas e táticas dependem de passos anteriores, como a análise de desafios, atividades desempenhadas pelo sistema, princípios, riscos, nível de impacto do modelo e outros (Hartmann Peixoto, 2020c, p. 154).

A implementação bem-sucedida dos princípios éticos da IA, para Zhou e Chen (2022, p. 7), depende da criação de um comitê de ética na organização e de uma colaboração estreita entre os criadores da ferramenta, os usuários e as pessoas afetadas pela utilização da IA. O comitê funcionaria como uma ponte entre a aplicação das diretrizes e as partes interessadas, a fim de tornar a ética da IA operável. Além disso, o comitê teria a função de confirmar se o sistema apresenta quaisquer riscos éticos, recomendar mudanças e até mesmo se posicionar contra o desenvolvimento ou aquisição de uma solução de IA (Zhou; Chen, 2022, p. 7-9).

Métricas qualitativas e quantitativas podem ser definidas para validar os princípios éticos da IA, tais medidas são estabelecidas a partir da importância para os usuários do sistema e da operabilidade para os desenvolvedores. Um padrão para as métricas de validação é útil para que as partes interessadas justifiquem a validação dos princípios na prática, mas deve-se levar em consideração que setores distintos têm diferentes ênfases nos princípios éticos. Assim, a ética da IA vai além de apenas identificar princípios éticos, mas também envolve construir padrões e colocá-los em prática. (Zhou; Chen, 2022, p. 10; p. 15).

A partir de uma revisão sistemática acerca das diretrizes e princípios para uma inteligência artificial ética, Jobin, Ienca e Vayena (2019, p. 394) observaram que a preservação e a promoção da justiça podem ser obtidas através de soluções técnicas como normas ou

codificação normativa explícita, fornecimento de informações e sensibilização do público para os direitos e regulamentações existentes, testes e auditorias, força de trabalho mais interdisciplinar e diversificada, bem como uma melhor inclusão da sociedade civil.

Construir ou manter uma IA fiável implica na adoção de processos para monitorar e avaliar a integridade dos sistemas ao longo do tempo e na adesão a ferramentas e técnicas que garantam a conformidade com normas e padrões. Algumas diretrizes exigem que a IA seja transparente ou explicável, a fim de se estabelecer a confiança, outras sugerem que esse princípio é atingido quando a IA cumpre as expectativas do público (Jobin; Ienca; Vayena, 2019, p. 395).

Ainda que as tecnologias de *machine learning* sejam caracterizadas como "sistemas fechados", a Estratégia Brasileira de IA (EBIA) dispõe que é possível incorporar a ideia de explicabilidade por meio da implementação de mecanismos para facilitar a rastreabilidade do processo decisório e do emprego de ferramentas e técnicas de interpretação, como auditabilidade e comunicação transparente sobre as capacidades do sistema (EBIA, 2021, p. 21).

É desejável que as etapas do processo de aprendizado de máquina que resultaram em uma inferência sejam rastreáveis e que as variáveis utilizadas passem por escrutínio, o que demonstra a importância dos processos de curadoria e seleção de dados a serem empregados. Ademais, a criação de rotinas de gestão de riscos, de monitoramento e de supervisão do sistema ao longo de todo o seu ciclo de vida contribuem para efetivação da interpretabilidade (EBIA, 2021, p. 24).

Ações estratégicas para a ética da IA indicadas pela EBIA incluem o estímulo do financiamento de projetos de pesquisa que visem aplicar soluções éticas principalmente nos campos de equidade/não-discriminação (*fairness*), responsabilidade/prestação de contas (*accountibility*) e transparência (*transparency*), conhecidas como matriz FAT (EBIA, 2021, p. 23).

É proposto pela EBIA que as licitações na administração pública tenham como requisito técnico o oferecimento, pelos proponentes, de soluções compatíveis com a promoção de uma IA ética. Portanto, estabelecer que soluções de tecnologia de reconhecimento facial adquiridas por órgãos públicos tenham um percentual de falso positivo abaixo de determinado limiar seria um exemplo dessa orientação (EBIA, 2021, p. 23).

A noção de *accountability* que advém da estratégia brasileira, traduzida pelo documento de referência como responsabilidade e prestação de contas, impõe que, a depender da aplicação de IA e dos riscos a ela associados, ocorra a implementação de mecanismos para a sua observância. Tais mecanismos incluem a designação de indivíduos ou grupos de uma organização para promover a conformidade com os princípios; treinamento para aumentar a conscientização interna sobre a necessidade de conformação; criação de certificações e códigos de conduta corporativos ou governamentais (EBIA, 2021, p. 24).

Ruster, Oliva-Altamirano e Daniell (2022, p. 3) apontam que mecanismos e ações para garantir um ecossistema que atenda ao princípio ético da dignidade incluem assegurar que os algoritmos atendam às leis antidiscriminação e a adoção de medidas para identificar e mitigar vieses nos conjuntos de dados. Os autores ressaltam a importância da criação de processos de feedback para que os usuários finais sejam ouvidos e suas necessidades sejam refletidas no design e monitoramento algorítmico, além da contratação de equipes com experiência relacionada ao local onde o sistema será utilizado.

Os modos de concretização do princípio da privacidade incluem: soluções técnicas, como privacidade desde a concepção; minimização de dados e controle de acesso; abordagens regulatórias, como a exigência de certificados ou a criação ou adaptação de leis e regulamentos para acomodar as especificidades da IA (Jobin; Ienca; Vayena, 2019, p. 395).

Segundo Arrieta et al. (2020, p. 108), a implementação dos princípios éticos da IA em uma organização requer a adoção de uma metodologia que inclua a conscientização e treinamento sobre possíveis problemas, tanto técnicos quanto não técnicos; ferramentas que ajudam a mitigar quaisquer problemas identificados; questionário que forneça orientações concretas sobre o que fazer se forem detectados determinados impactos indesejados; e, um modelo de governança que atribua responsabilidades e prestações de contas.

Para que haja a correta utilização de métodos de IA em organizações e instituições, é essencial que os princípios éticos aplicáveis sejam estudados conjuntamente. Ademais, o cumprimento desses princípios na implementação de modelos na prática é um caminho para se chegar em direção a uma IA responsável, indicam Arrieta et al. (2020, p. 108).

A variedade de documentos orientadores atesta o interesse crescente na ética da IA por parte da comunidade internacional e por diferentes tipos de organização. A proporção quase equivalente de diretrizes emitidas pelo setor público e pelo setor privado sugere que a ética da IA diz respeito às duas esferas. Contudo, existem áreas globais sub-representadas, como África,

América do Sul e Central e Ásia Central, o que indica que as regiões globais não participam de forma igualitária desse debate, revelando um desequilíbrio de poder no discurso internacional (Jobin; Ienca; Vayena, 2019, p. 396).

Cozman e Kaufman (2022, p. 209) alertam que o uso de sistemas de IA em razão da "promessa de objetividade", com a suposição de que os algoritmos garantem objetividade e/ou neutralidade por serem processados por máquinas e protegidos dos erros humanos é um desacerto. Recomenda-se que os modelos de IA sejam vistos como parceiros dos especialistas humanos e a ética da IA deve ter sua compreensão voltada a mitigar riscos e eleger prioridades, ante a impossibilidade de se controlar todos os desenvolvimentos e usos.

O Painel de Projetos de IA no Poder Judiciário - 2023, elaborado no âmbito do Programa Justiça 4.0, elencou as respostas fornecidas pelos tribunais sobre as medidas adotadas ou que se pretende adotar em relação à transparência e ética no uso de IA (CNJ, 2023). Na tabela abaixo foram reproduzidas as principais respostas dadas:

Tabela 3 - Principais medidas que se pretende adotar/adotadas pelos tribunais em relação à transparência e ética no uso de IA

Medidas que se pretende adotar/adotadas em relação à transparência e ética no uso de IA nos tribunais	**Quantidade de respostas**
Fomentar informações aos usuários sobre a existência de IA.	102
Treinamento e capacitação dos servidores e magistrados sobre o uso ético de IA.	90
Estabelecimento de diretrizes e políticas claras sobre o uso de IA no judiciário.	74
Implementação de mecanismos de explicabilidade dos resultados obtidos pela IA.	56
Realização de auditorias periódicas nos algoritmos e modelos de IA utilizados.	38
Disponibilização pública do código-fonte e funcionamento dos sistemas de IA.	25
Não sei	21
Existe proposta de criação de área específica STI para estudo e implantação de projetos relacionados às preocupações relacionadas à ética e transparência	5

Disponibilização dos modelos e *datasets* de treinamento no Sinapses	3

Tabela elaborada com base nos dados do Painel de Projetos de IA no Poder Judiciário - 2023 (CNJ, 2023).

Siqueira, Morais e Santos (2022, p. 12) consideram que não é mais possível vislumbrar a atividade jurisdicional de maneira seccionada dos avanços tecnológicos. Diante disso, os autores afirmam que uma medida relacionada à transparência que poderia ser adotada é a indicação, no próprio sistema de processo eletrônico, de que determinada decisão foi baseada em ferramenta de IA, como por exemplo: "Esta decisão foi exarada com auxílio de mecanismos de inteligência artificial, nos termos da Res. n. 332 CNJ" (Siqueira; Morais; Santos, 2022, p. 24).

Em vista da crescente implementação de soluções baseadas em IA, Salomão e Tauk (2023b, p. 28-29) sugerem que os tribunais adotem três medidas, levando em consideração a Resolução CNJ n. 332/2020. A primeira delas consiste em mapear e avaliar os resultados, quantitativos e qualitativos, do uso dos sistemas em comparação com a situação anterior à respectiva utilização. A segunda medida enfatiza a transparência, com o fornecimento de explicações sobre os sistemas de IA nas páginas eletrônicas dos tribunais, noticiando suas tarefas e finalidades, de modo acessível e compreensível ao público externo. A terceira providência versa sobre o controle prévio do treinamento do modelo computacional, em atenção aos dados utilizados e com a formação das equipes para pesquisa, desenvolvimento e implantação, já que o controle posterior pode não ser efetivo (Salomão; Tauk, 2023b, p. 28-29).

4 PROJETOS DE PESQUISA E DESENVOLVIMENTO (P&D) E A INTELIGÊNCIA ARTIFICIAL

De acordo com Nonato (2023), Pesquisa e Desenvolvimento (P&D) é o processo dedicado a criar algo ou buscar o aperfeiçoamento do que já existe em uma organização. A pesquisa pode ser definida como o esforço para obter informações relevantes, enquanto o desenvolvimento é a utilização de recursos para a criação e ampliação dos seus resultados. Quanto se junta os dois aspectos, Pesquisa e Desenvolvimento funcionam como um ciclo de aperfeiçoamento, seja de maneira contínua ou transformacional (Nonato, 2023).

Na base de conhecimento da Organização para Cooperação e Desenvolvimento Econômico (OCDE), indica-se que termo Pesquisa e Desenvolvimento (P&D) compreende o trabalho criativo realizado de forma sistemática para aumentar o estoque de conhecimento humano e conceber novas aplicações baseadas nele. Além disso, é estabelecido que a Pesquisa e Desenvolvimento abrange as atividades de pesquisa básica, pesquisa aplicada e desenvolvimento experimental (OCDE, 2024b).

A pesquisa básica é o trabalho experimental ou teórico realizado principalmente para adquirir novos conhecimentos, sem nenhuma aplicação ou uso específico em vista. Na pesquisa aplicada também ocorre uma investigação original para aquisição de novos conhecimentos, mas há o direcionamento para um objetivo ou meta prática específica. O desenvolvimento experimental é o trabalho sistemático, que aproveita o conhecimento obtido de pesquisa e/ou experiência prática, para produzir novos materiais, produtos ou dispositivos, ou para melhorar aqueles já produzidos ou instalados (OCDE, 2024b).

Segundo Pinheiro et al. (2006, p. 460), os projetos de P&D convivem com um componente de incerteza com relação aos seus resultados. Quanto maior o desconhecimento do resultado esperado, maior é o risco relacionado ao projeto. Em uma proposta de metodologia dividida em etapas, o projeto é reavaliado para decidir-se por sua continuidade ou não ao final de cada fase. Para os autores, a pesquisa, na maioria das organizações, está baseada em uma estrutura acadêmica, disciplinar, com alto grau de especificidade, enquanto o desenvolvimento tecnológico é multidisciplinar e deve focar o mercado (Pinheiro et al., 2006, p. 460).

A diversidade de especialidades profissionais necessárias para a realização dos projetos, o número de pessoas envolvidas, as instalações físicas necessárias, o volume de informações a serem processadas e rastreadas, a real duração do projeto e o número de parceiros envolvidos

para o desenvolvimento e conclusão são fatores que demonstram a complexidade dos projetos de Pesquisa e Desenvolvimento e apontam para a necessidade de um assíduo acompanhamento e controle das atividades (Pinheiro et al., 2006, p. 461).

Pinheiro et al. (2006, p. 460) propõem uma metodologia para servir de orientação aos projetos de Pesquisa e Desenvolvimento, consistente em um conjunto de etapas que devem ser adaptadas à realidade da organização:

Tabela 4 - Síntese de metodologia para projetos de Pesquisa e Desenvolvimento

Metodologia para projetos de Pesquisa e Desenvolvimento (P&D)	
Macroetapas	Detalhamento específico
1. Diagnóstico da situação	• Definição do ciclo de vida do projeto a ser desenvolvido;
	•Identificação das lacunas na disponibilidade de recursos humanos;
	• Identificação do staff [com conhecimento] técnico requerido
	• Reconhecimento das limitações gerenciais/operacionais, considerando-se a infraestrutura disponível, a estrutura organizacional e a missão institucional.
2. Identificação das características do processo que possam gerar impacto em fatores de risco, custo e tempo	• Projetos multidisciplinares x projetos concentrados em uma área específica;
	• Projetos caracterizados com elevado grau de incerteza com relação aos resultados esperados e elevada complexidade, em função da multidisciplinaridade;
	• Desconhecimento das regulamentações gerais (legislação) e específicas (guidelines).
3. Formulação da situação desejada	• Adequação dos projetos de pesquisa para a obtenção de produtos;
	• Desenvolvimento e adaptação de métodos para facilitar a gestão dos projetos;
	• Criação de estrutura de suporte técnico-gerencial.

4. Objetivos	• Implantação e realização de estudos periódicos de viabilidade técnico-econômica (EVT) para os projetos, dentro de um processo gerenciado formalmente, e em conjunto com a equipe multidisciplinar;
	• Garantia de suporte técnico-gerencial aos pesquisadores;
	• Integração paulatina das áreas multidisciplinares no sentido de torná-las interdisciplinares;
5. Elaboração da proposta de gestão	• Criação e estabelecimento de uma equipe multidisciplinar, contemplando novos cargos e funções, adequados à execução do projeto e ao gerenciamento das suas fases técnicas específicas;
	• Implantação de um sistema documental.
6. Instrumentalização da gestão para o acompanhamento, avaliação e controle dos projetos	• Elaboração do escopo dos projetos, definindo todos os passos técnicos, científicos e gerenciais desde a fase inicial da pesquisa até o desenvolvimento do produto, de forma a otimizar tempo e recursos, respaldados na regulamentação da área;
	• Utilização de ferramentas e sistemas informatizados de acompanhamento de projetos.

Fonte: Pinheiro et al., 2006 (com adaptações).

No mundo todo, o investimento em Pesquisa e Desenvolvimento (P&D) relativo a projetos em IA cresceu significativamente, realizados tanto por empresas como por governos. Em 2020, por exemplo, o investimento do governo dos EUA em P&D nessa área foi de aproximadamente US$1,5 bilhões de dólares. (Littman et al., 2021, p. 39).

Segundo o Índice Global de Inovação 2023, o Brasil ocupa a 49ª posição em um total de 132 economias no quesito inovação. Em 2022, o país investiu em Pesquisa e Desenvolvimento (P&D) o equivalente a 1,2% do seu PIB. As três primeiras economias do ranking, Suíça, Suécia e Estados Unidos, investiram respectivamente 3,2%, 3,3% e 3,5% do Produto Interno Produto em Pesquisa e Desenvolvimento (WIPO, 2023, p. 96; p. 193; p. 194; p. 206). A partir da análise dos dados mencionados, é possível refletir que a expansão dos investimentos em P&D pelo Brasil poderia propiciar um melhor ambiente de inovação no país.

No âmbito da OCDE, entre as recomendações para os gestores de políticas públicas está o investimento na Pesquisa e Desenvolvimento em IA. Segundo a recomendação, os governos

devem considerar o investimento público de longo prazo e incentivar o investimento privado em Pesquisa e Desenvolvimento e *open science*, incluindo esforços interdisciplinares para estimular a inovação em IA confiável, que se concentre em questões técnicas desafiadoras e em implicações sociais, legais e éticas (OCDE, 2024).

É também aconselhado, pela OCDE, o incentivo ao investimento em ferramentas de código aberto e em conjuntos de dados abertos que sejam representativos e respeitem a privacidade e a proteção de dados, de forma a dar suporte a um ambiente de Pesquisa e Desenvolvimento em IA livre de vieses prejudiciais e para melhorar a interoperabilidade e o uso de padrões (OCDE, 2024).

A Estratégia Brasileira de Inteligência Artificial (EBIA) indica ações estratégicas para pesquisa, desenvolvimento, inovação e empreendedorismo na área da inteligência artificial, as quais estão sistematizadas na tabela abaixo:

Tabela 5 - Ações estratégicas para pesquisa, desenvolvimento, inovação e empreendedorismo na área da IA

Ações Estratégicas
∗ Definir áreas prioritárias para investimentos em IA, de maneira alinhada a outras políticas relacionadas ao ambiente digital.
*Ampliar as possibilidades de pesquisa, desenvolvimento, inovação e aplicação de IA, por meio da viabilização do aporte de recursos específicos para esse tema e da coordenação entre iniciativas já existentes.
* Estabelecer conexões e parcerias entre setor público, setor privado e instituições científicas e universidades em prol do avanço no desenvolvimento e utilização da IA no Brasil.
* Promover um ambiente de políticas públicas que apoie uma transição ágil da fase de P&D para a fase de desenvolvimento e operação de sistemas de IA
* Promover um ambiente para Pesquisa e Desenvolvimento em IA que seja livre de viés
* Aperfeiçoar a interoperabilidade e o uso de padrões comuns.
* Promover mecanismos de incentivo que estimulem o desenvolvimento de sistemas de IA que adotem princípios e valores éticos.

Fonte: EBIA, 2021, p. 37.

Segundo Peter Cihon (2019, p. 31-32), os padrões (*standards*) podem ser utilizados para propagar uma cultura de segurança e responsabilidade na Pesquisa e Desenvolvimento em IA. Esse objetivo pode ser atingido de 4 formas; na primeira, os critérios descritos como padrão em uma organização definem orientações e expectativas. Ao adotar um padrão de transparência, por exemplo, a organização compromete-se com esse propósito nos sistemas de IA. A segunda maneira seria por meio da adoção de normas internacionais, a terceira é internalizar na rotina práticas de responsabilidade e segurança, e a quarta maneira consiste em incorporar os *standards* diretamente nos produtos e *softwares* desenvolvidos.

Para Hartmann Peixoto (2020c, p. 8), a preocupação com parâmetros éticos está no próprio fundamento da Pesquisa e Desenvolvimento de aplicações de IA para o Direito, principalmente as que se destinam aos fluxos de gestão processual e o apoio à decisão, diretamente relacionadas à concretização de direitos. Para o autor, sem os referenciais éticos há o risco de se fornecer elementos à justificação de mitos associados à IA, como a frieza e o preconceito; e também de se esvaziar o conteúdo positivo de uma possível aplicação.

A formalização das necessidades éticas em um projeto de P&D possibilita uma reflexão acerca das melhores formas de comunicação e esclarecimento sobre os objetivos e impactos da proposta. Além disso, a partir de um projeto ético viabiliza-se a criação de um sistema de controle, tanto para a aferição de benefícios quanto para a identificação e gestão de danos (Hartmann Peixoto; Silva, 2019, p. 40).

No âmbito das recomendações táticas para implementação da ética em um sistema de IA, tem-se que Pesquisa e Desenvolvimento (P&D) de uma ferramenta deve deixar claro quais são as diretrizes éticas para o tratamento do sistema cognitivo artificial e deve conter perguntas que possam ser usadas, de uma forma geral, para determinar deveres em contextos sociais (Hartmann Peixoto, 2020c, p. 154).

De acordo com a EBIA, a Pesquisa e Desenvolvimento da IA deve adotar abordagens éticas de design para tornar o sistema confiável. Isso inclui, mas não se limita a "tornar o sistema o mais justo possível, reduzir possíveis discriminações e preconceitos, melhorar sua transparência, prover explicação e previsibilidade e tornar o sistema mais rastreável, auditável e responsável" (EBIA, 2021, p. 36).

4.1 Contexto e objetivos do Projeto de Pesquisa e Desenvolvimento (P&D) Sabiá

Fruto de uma parceria entre o Tribunal Superior do Trabalho e a Universidade de Brasília, por meio dos Laboratórios de Pesquisa AI.Lab[7] e DR.IA[8], e com o intermédio da Fundação de Empreendimentos Científicos e Tecnológicos (Finatec), o Projeto Sabiá - Processamento de Linguagem Natural Aplicado ao Sistema Bem-Te-Vi para Análises de Processos Jurídicos do Tribunal Superior do Trabalho foi iniciado em 2022 e, até o fechamento do presente trabalho, permanecia em desenvolvimento.

Para uma melhor compreensão das finalidades deste Projeto de Pesquisa e Desenvolvimento (P&D) é relevante entender o âmbito de sua inserção: a Justiça do Trabalho. Mantida pela União, esse ramo especializado da Justiça tem sua competência indicada no art. 114 da Constituição Federal, sendo responsável por conciliar e julgar as ações judiciais oriundas da relação de trabalho (abrangidos os entes de direito público externo e da administração pública direta e indireta da União, dos Estados, do Distrito Federal e dos Municípios), as que envolvam o exercício do direito de greve, as ações sobre representação sindical, além das demandas que tenham origem no cumprimento de suas próprias sentenças (CNJ, 2024, p. 40).

São órgãos da Justiça do Trabalho: o Tribunal Superior do Trabalho (TST), os 24 Tribunais Regionais do Trabalho (TRTs) e os juízes do trabalho atuantes nas varas do trabalho. Nas comarcas não abrangidas pela jurisdição da Justiça do Trabalho, a competência será atribuída aos juízes de Direito, com recurso para o respectivo Tribunal Regional do Trabalho (CNJ, 2024, p. 40).

O Tribunal Superior do Trabalho (TST) é o órgão máximo da Justiça do Trabalho e tem como principal função uniformizar as decisões sobre ações trabalhistas, consolidando a jurisprudência desse ramo do Direito (CNJ, 2024, p. 46). Esse Tribunal Superior é composto por 27 ministros (as) e apresenta a seguinte estrutura interna:

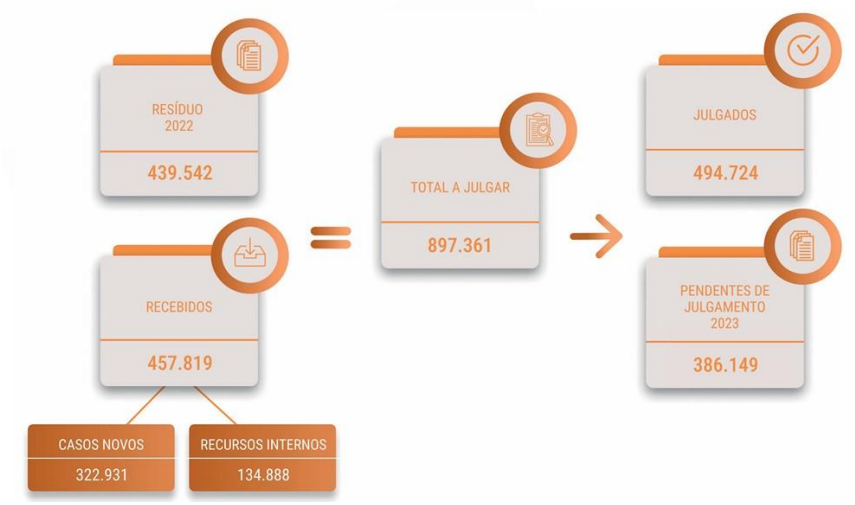
[7] https://ailab.unb.br/
[8] http://dria.unb.br/

Figura 16 - Estrutura do Tribunal Superior do Trabalho

Fonte: TST, 2024.

Em 2023, o TST recebeu 457.819 processos para julgamento, dos quais 322.931 eram casos novos e 134.888 representavam recursos internos. O resíduo de 2022, que totalizava 439.542 processos, ampliou o acervo para 897.631processos a julgar em 2023. Foram julgados 494.724 processos e 386.149 processos ficaram pendentes de julgamento (TST, 2024, p. 141-151). A figura a seguir sintetiza os dados do acervo:

Figura 17 - Resumo do acervo do TST em 2023

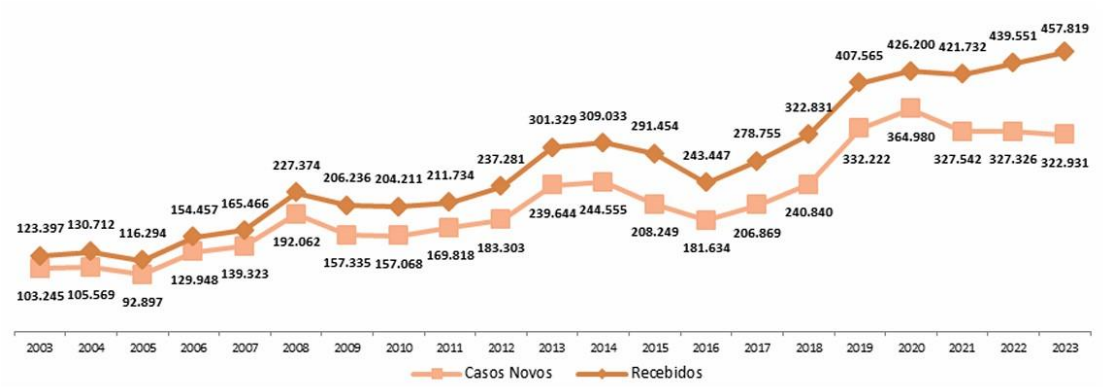


Fonte: TST, 2024

Um volume expressivo dos processos que ingressam no Tribunal corresponde às classes processuais Recurso de Revista (RR), Agravo de Instrumento em Recurso de Revista (AIRR) e Recurso de Revista com Agravo (RRAg). A título de exemplo, em 2023 foram recebidos 248.642 Agravos de Instrumentos em Recurso de Revista, e apesar do marcante número, um

total de 305.061 processos desta classe foram julgados, o que significa que parte do resíduo foi absorvido. Ainda que a produtividade do Tribunal seja elevada, a série histórica retrata uma tendência crescente de casos recebidos.

Figura 18 - Série histórica de recebidos e de casos novos de 2003 a 2023



Fonte: TST, 2024

O cenário de grande litigiosidade enfrentado pelo TST e as peculiaridades dos recursos dirigidos à Corte, uma vez que uma mesma peça recursal pode conter diversos temas que precisam ser analisados separadamente, como se fossem recursos distintos, é favorável à otimização dos esforços por meio da utilização de inteligência artificial. As soluções de IA, portanto, podem ser pensadas para enfrentar os desafios na gestão do acervo e proporcionar uma melhor prestação jurisdicional.

Nessa perspectiva, foi idealizado o Projeto Sabiá, que tem como escopo pesquisar e desenvolver soluções para complementar o sistema Bem-Te-Vi (sistema interno do TST) no que se refere às funcionalidades de agrupamento de processos por similaridade (Módulo iSimilares) e levantamento de jurisprudência (Módulo iJulgados).

O Bem-Te-Vi é um sistema de gerenciamento de processos que utiliza tecnologias de *big data* para disponibilizar aos gabinetes do TST informações sobre os processos de seu acervo, para apoio à gestão e triagem. O software utiliza inteligência artificial para fazer a análise automática da tempestividade (observância dos prazos) dos processos e começou a funcionar em outubro de 2018, viabilizando a definição de estratégias para aumento da produção dos gabinetes (CSJT, [s.d]).

Observa-se que tanto o sistema Bem-Te-Vi quanto o Sabiá foram contabilizados no Painel da Pesquisa sobre Inteligência Artificial 2023, elaborado pelo Conselho Nacional de Justiça. O primeiro foi classificado como “finalizado e em produção” e o segundo foi categorizado como “em andamento” (CNJ, 2023).

O Módulo iSimilares do Projeto Sabiá consiste na Pesquisa e Desenvolvimento (P&D) de uma solução capaz de selecionar um subconjunto de processos similares a um processo de referência. A atuação do software ocorre sobre a peça Despacho de Admissibilidade, de modo a segmentar os temas identificados na referida peça. Para cada tema é feita uma classificação de processos similares em forma de ranking, em que processos mais similares em relação ao processo de referência apresentam um maior percentual de semelhança indicado pela ferramenta. Esse primeiro módulo tem o objetivo de proporcionar maior agilidade e assertividade na gestão do acervo processual, permitindo que o usuário realize o agrupamento de processos similares sem a utilização prévia de filtros.

De acordo com Lemos, Torres e Rocha (2024, p. 8-9), a escolha do Despacho de Admissibilidade deu-se após uma análise quantitativa e qualitativa dos tipos de peças mais relevantes para se encontrar a similaridade entre os processos. Nesse tipo de decisão consta a análise tema a tema feita pelos Tribunais Regionais do Trabalho a respeito do cabimento do Recurso de Revista dirigido ao TST. Essa peça é frequentemente redigida por meio da ferramenta eRec (módulo do PJe), que guarda informações estruturadas de cada tema do processo.

Por meio dos modelos Bert1 e K-means2, o texto de Despacho de Admissibilidade é processado e então agrupado para cada assunto, permitindo o encontro de nuances de discussões dentro de um mesmo assunto e a indicação dos processos mais similares na temática. Quando o eRec não é utilizado na origem, verificou-se uma maior dificuldade de separação das informações de cada tema e também a impossibilidade de relacionar os assuntos à Tabela Unificada de Assuntos do CNJ, uma vez que não havia uniformidade na redação dos mesmos. Assim, documentos não originários do eRec deixaram de ser escopo do Projeto Sabiá (Lemos; Torres; Rocha, 2024, p. 8-9).

Para os documentos provenientes do eRec, Lemos, Torres e Rocha (2024, p. 9) apontam que os grupos formados para cada assunto apresentaram uma mediana de 95% de similaridade. De acordo com as pesquisadoras, após a avaliação pela equipe negocial, os resultados

preliminares foram favoráveis e indicam que as similaridades sugeridas para um mesmo tema estão apropriadas.

O Módulo iJulgados, por sua vez, consiste na Pesquisa e Desenvolvimento (P&D) de uma solução que auxilie os gabinetes na recuperação de julgados (decisões monocráticas e acórdãos) emitidos pelo gabinete do usuário, turma do usuário, outras turmas, subseção e órgão especial do TST. Para isso, a atuação do robô ocorre nas peças do Despacho de Admissibilidade, do Recurso de Revista e do Agravo de Instrumento em Recurso de Revista, buscando correlacionar processos que já foram julgados a um determinado processo de referência.

Ressalta-se que esse sistema de IA não confecciona peças e não aponta que determinado julgado é aplicável ao processo em análise. A ferramenta indica percentuais de similaridade que devem ser verificados pelo usuário, o qual deve ponderar sobre o mais adequado encaminhamento para a questão em debate.

Segundo Siqueira, Silva e Correa (2024, p. 4-5), o Projeto Sabiá envolve a pesquisa de métodos de aprendizado de máquina (AM) para enfrentar o desafio da busca por similaridade. Para os autores, a aplicação do modelo Sentence-BERT, técnica de aprendizado de máquina baseada em transformadores que permite capturar a semântica de sentenças textuais e medir a similaridade entre elas, demonstrou eficácia na identificação de sentenças semelhantes em peças de Recursos de Revista (RR), diante do pré-treinamento da ferramenta com dados jurídicos.

Para além dos dois módulos anteriormente descritos, o projeto visa pesquisar recursos de *Long Life Machine Learning* (LLML) com o objetivo de desenvolver uma solução que traga um ciclo de vida longo aos modelos de IA implantados e também um sistema de registro de *feedbacks* do usuário, a fim de evidenciar o desempenho real das ferramentas (CNJ, 2023).

Diversamente dos *softwares* de prateleira, caracterizados pela utilização em larga escala, ausência de recursos ou funcionalidades personalizadas e o atendimento genérico a diversos consumidores (Shono, 2023), o Projeto Sabiá é desenvolvido com base nas necessidades e expectativas do Tribunal Superior do Trabalho. Para isso, o Projeto possui uma equipe jurídica e uma equipe da área da tecnologia em constante interação e que dialogam em todas as etapas e avaliações com a equipe negocial do TST.

Com a implantação do Projeto de Pesquisa e Desenvolvimento (P&D) Sabiá, espera-se obter a otimização dos recursos humanos envolvidos na condução das atividades

administrativas e judiciais do Tribunal Superior do Trabalho; o aumento da capacidade de processamento do volume de demandas e a diminuição da taxa de congestionamento de processos no Tribunal. Sob a perspectiva do jurisdicionado, este poderá ser beneficiado com decisões mais céleres, com a consequente redução do tempo de tramitação do seu processo, e também com uma tutela mais justa, minimizando-se o risco de decisões divergentes para casos semelhantes que deveriam ser tratados de forma similar.

4.2 Análise do Projeto de Pesquisa e Desenvolvimento (P&D) Sabiá à luz da transparência e ética da IA

O Projeto Sabiá está em fase de desenvolvimento e ainda não foi implantado no Tribunal Superior do Trabalho. Apesar disso, é possível analisar aspectos éticos e requisitos de transparência que estão presentes nessa fase da Pesquisa e Desenvolvimento e também fazer considerações sobre os critérios éticos que podem ser implementados.

No Painel da Pesquisa sobre Inteligência Artificial 2023, transmitido pelo Conselho Nacional de Justiça, foi divulgada uma lista de perguntas elaboradas pelo referido Conselho e as respectivas respostas dadas pelos tribunais. No caso do Projeto Sabiá, algumas das respostas fornecidas pelo TST à pesquisa foram reproduzidas abaixo:

Tabela 6 - Perguntas e Respostas sobre o Projeto Sabiá

Perguntas e Respostas sobre o Projeto Sabiá	
Perguntas	**Respostas**
Nome do Projeto	Sabiá
Descrição do Projeto	O projeto tem como escopo pesquisar e desenvolver soluções para complementar o sistema Bem-Te-Vi no que se refere às funcionalidades de Agrupamento de Processos e Levantamento de Jurisprudência. Visando evidenciar o desempenho real destas soluções, é objetivo deste trabalho desenvolver um módulo de Registro de Feedbacks dos usuários. Por fim, este projeto também visa pesquisar soluções de Long Life Machine Learning (LLML) com o objetivo de desenvolver uma solução que traga um ciclo de vida longo aos modelos de IA implantados. Em síntese, este projeto visa o desenvolvimento de 2 (duas) soluções de IA, um sistema de registro de desempenho destas IAs e uma solução de LLML, conforme especificado abaixo:

	1. Agrupar Processos: P&D de uma solução capaz de selecionar um subconjunto de processos similares a um processo de referência; 2. Identificar Jurisprudência : P&D de uma solução capaz de selecionar um subconjunto de decisões e acórdãos da base de jurisprudência que tenha correlação com um processo de referência.
Qual é o estágio de evolução desse projeto?	Em andamento.
Qual a infraestrutura disponível no Tribunal/Conselho para desenvolver e/ou incorporar esse projeto?	Possui parque computacional somente de CPUs.
Como esse projeto foi desenvolvido por seu Tribunal/Conselho no que diz respeito a colaboração com outras entidades?	Em parceria com universidades.
Qual é o tamanho da equipe envolvida no projeto?	5
Qual a origem dos dados usados para o treinamento do(s) modelo(s)?	Gerados pelo próprio tribunal; obtidos através de diferentes órgãos do judiciário.
Quais foram os principais resultados e benefícios alcançados com a adoção desse projeto de IA no seu Tribunal/Conselho?	Identificação de padrões e tendências em grandes volumes de dados jurídicos; maior eficiência e agilidade no processamento de documentos e informações; otimização de recursos e redução de custos operacionais; redução do tempo de tramitação dos processos judiciais.
Seu Tribunal/Conselho tem acesso ao código-fonte dos algoritmos usados no projeto?	Sim
Seu Tribunal/Conselho possui(u) uma equipe dedicada para esse projeto?	Sim.

Seu Tribunal/Conselho planeja utilizar "LLMs - Large Language Models" em suas soluções de IA para apoiar em atividades administrativas? E em atividades jurisdicionais?	Não, ainda não utilizamos "LLMs", mas temos planos de explorar essa tecnologia no futuro.
Seu projeto usa/utilizará aprendizado de máquina (Machine Learning)?	Sim.
Quais são as principais preocupações éticas relacionadas ao uso de IA no seu Tribunal/Conselho?	Qualidades das informações produzidas por IA.
Quais medidas seu Tribunal/Conselho adota ou pretende adotar para garantir a transparência e ética no uso de IA no Poder Judiciário?	Implementação de mecanismos de explicabilidade dos resultados obtidos pela IA.
Quais frameworks foram/serão utilizados nesse projeto?	Scikit learn.
Assinale abaixo as opções que melhor identificam a(s) tarefa(s) com base no algoritmo:	Kmeans; dbscan.
Assinale abaixo as opções que melhor identificam a(s) tarefa(s) utilizada(s) no projeto:	Similaridade de texto; sumarização.
Assinale abaixo as opções que melhor identificam a(s) atividade(s) contempladas por esse projeto:	Busca de casos similares.

A quem pertence o algoritmo utilizado no projeto?	Código aberto.
Os dados e resultados gerados por seu projeto passaram por :	Monitoramento técnico durante todo o processo de desenvolvimento visando garantia de qualidade;
O projeto possui documentação?	Sim.
O código-fonte do(s) algoritmo(s) do projeto está/estará disponível publicamente para reutilização?	Sim.
Assinale abaixo as opções que melhor classifica o(s) métodos(s) de aprendizado utilizado(s) no projeto:	Aprendizado Não Supervisionado.
A solução desenvolvida no projeto é/será consumida de que forma?	Através de uma implementação local.

Fonte: CNJ, 2023.

As perguntas e respostas transcritas são de acesso público e podem ser localizadas via painel interativo do CNJ, na fonte citada acima. Observa-se o esforço comunicativo do TST ao transmitir, mesmo antes do lançamento, diversas informações sobre o desenvolvimento do Projeto Sabiá. No material divulgado, elucidou-se a criação de duas soluções de IA para complementar o sistema Bem-Te-Vi, contemplando a busca de casos similares, o registro de feedbacks dos usuários e o objetivo de implantar um modelo com um ciclo de vida longo; demonstrou-se a intenção de uso do aprendizado de máquina não supervisionado e a divulgação do código fonte para reutilização; foi explicado que ocorre o monitoramento técnico durante todo o processo de desenvolvimento visando a garantia de qualidade. Ademais, o Tribunal manifestou a pretensão de explorar os *Large Language Models* no futuro, os quais estão diretamente relacionados ao subcampo da IA generativa.

O princípio ético da explicabilidade, que denota ações ou procedimentos realizados com a intenção de esclarecer ou detalhar as funções internas de um modelo (Arrieta et al., 2020, p. 84), pode ser notado na descrição do projeto, em que são elencados os objetivos e as

funcionalidades do modelo, assim como na indicação de que serão feitas tarefas de busca por similaridade de texto e de busca de processos similares, o que permite a um observador externo compreender as atividades que serão exercidas pelo sistema. Após a implantação, uma divulgação sobre o modo como a ferramenta alcança determinados resultados a partir dos dados escolhidos para a entrada contribuiria para uma melhor verificação desse princípio pelos usuários e afetados.

Conforme exposto no terceiro capítulo, para Ryan e Stahl (2021, p. 66) o princípio da transparência pode ser observado pela transparência da própria tecnologia de IA e a transparência das organizações de IA que a desenvolvem e utilizam-na. Embora o modelo de IA do Projeto Sabiá ainda esteja em construção, o TST anunciou a origem dos dados para o treinamento, as tarefas dos algoritmos de aprendizado de máquina que se pretende utilizar, o emprego do framework Scikit learn (biblioteca de código aberto), o método de aprendizado máquina escolhido para a ferramenta, a forma como será implementada a solução e expressou-se a intenção de disponibilizar publicamente o código fonte para reutilização.

Do ponto de vista tecnológico, nota-se a existência da transparência ao serem noticiadas as características da tecnologia por trás da ferramenta em desenvolvimento. Ademais, detecta-se a transparência, do ponto de vista da entidade que utilizará o modelo, quando o Tribunal informa os possíveis benefícios que podem advir para a prestação jurisdicional, mas sem deixar de sinalizar a preocupação ética acerca da qualidade das informações produzidas por IA.

Expressou-se no questionário que o Projeto está sendo desenvolvido em colaboração com universidade. Parcerias entre tribunais e a academia podem resultar em uma desejável diversidade das equipes, conforme preconizado pelo art. 20 e parágrafos da Resolução CNJ n. 332/2020. Contudo, não foram especificadas, nesse primeiro momento, informações sobre a quantidade de pesquisadores externos e as áreas de conhecimento dos profissionais envolvidos, de modo que tal indicação contribuiria para uma maior transparência.

Há o monitoramento técnico durante todo o processo de desenvolvimento da ferramenta visando a garantia de qualidade, de maneira que a equipe interna do Tribunal participa de cada fase do Projeto. Esse acompanhamento constante é compatível com o princípio ético da proteção, segundo o qual todas as etapas de desenvolvimento e uso da IA devem estar preenchidas de mecanismo de segurança e monitoramento (Hartmann Peixoto, 2020c, p. 147). Ressalta-se que é importante que a supervisão continue mesmo após a implantação do modelo, a fim de identificar a adequação do seu funcionamento.

O princípio da justiça substancial, na percepção de Hartmann Peixoto (2020c, p. 143), significa que os sistemas de IA têm a responsabilidade ativa pela realização de justiça e o compromisso com a inclusão e equidade. Nesse sentido, considerando o uso da ferramenta no apoio à atividade jurisdicional, o princípio da justiça substancial pode ser verificado ao se buscar uma melhor gestão do acervo, para que casos similares sejam tratados de maneira semelhante e na redução do tempo de tramitação dos processos judiciais.

Os sistemas de IA implantados no Poder Judiciário não utilizam, como regra, dados sensíveis (Salomão; Tauk, 2023, p.77). O Projeto Sabiá, pela descrição das suas finalidades e objetivos encaixa-se nessa regra de não utilização desse tipo de dado. O princípio da privacidade dos dados está ligado ao manejo adequado e à salvaguarda de dados privados e públicos, dotados de algum grau de sensibilidade (Hartmann Peixoto, 2020c, p. 147). No Projeto em questão, foi informado que os dados usados para o treinamento do modelo provêm do próprio Tribunal e são obtidos através de diferentes órgãos do judiciário, contudo, não se indica que tipo de dados são esses. Assim, considerando o princípio ético da privacidade dos dados, propõe-se que nas futuras divulgações de resultados seja comunicado quais tipos de dados são utilizados e como se dá o manejo destes.

Segundo Salomão e Tauk (2023, p. 79-80), a maioria dos tribunais costuma divulgar, tão somente, os objetivos e as aplicações da ferramenta, sem uma consolidação dos resultados e da evolução destes ao longo do tempo e, tampouco, divulga a documentação da ferramenta de IA. Nesse sentido, sugere-se que após a implantação do Projeto Sabiá ocorra a publicização da documentação do Projeto à coletividade e que o próprio Tribunal se comprometa com anúncios periódicos acerca dos resultados obtidos, o que tornaria possível a aferição do princípio da *accountability*.

A comunicação sobre o uso das ferramentas de IA no judiciário está disponível, em geral, como notícia nos sítios eletrônicos dos tribunais. Questiona-se se essa disponibilização cumpre de forma adequada ao requisito de aviso aos usuários externos do uso de IA nos serviços que lhe são prestados, uma vez que essas informações costumam não ser conhecidas, inclusive, pelos próprios advogados (Salomão e Tauk, 2023, p. 84). Desse modo, no âmbito do Projeto Sabiá, propõe-se que além de uma comunicação no site do TST, após a implantação da ferramenta, seja dado o aviso no próprio sistema de processo eletrônico de que há o uso de inteligência artificial como apoio às decisões judiciais, conforme defendem Siqueira, Morais e Santos (2022, p. 24).

Os princípios da ética da IA são dinâmicos e a verificação destes depende do contexto em que a inteligência artificial será utilizada e das atividades que serão executadas pelos sistemas. Assim, na presente seção realizou-se uma escolha dos princípios éticos que seriam mais relevantes para se analisar um Projeto de Pesquisa e Desenvolvimento em IA direcionado ao Poder Judiciário. O princípio da transparência certamente prepondera nesse debate e possui interseções com os princípios da explicabilidade, justiça substancial, privacidade dos dados e *accountability*, os quais também foram selecionados.

Constata-se que o conjunto de informações disponibilizadas pelo Tribunal Superior do Trabalho via CNJ são necessárias, mas não suficientes ao adequado cumprimento de alguns dos princípios éticos analisados aqui. Por esse motivo, foram apresentadas uma série de proposições, que objetivam ser um exercício crítico e ao mesmo tempo direcionador para uma futura divulgação do Projeto Sabiá.

CONCLUSÃO

Destaca-se que apesar de terem sido encontradas diferentes definições para o termo inteligência artificial, existem intersecções entre elas, como a noção de resolução de problemas específicos, a tentativa de aproximação com a cognição humana e o entendimento de que a área é permeada pela interdisciplinaridade. Ademais, os conceitos não são estanques e podem passar por atualizações a fim de incluir ideias relacionadas às evoluções dos modelos, a exemplo da mudança promovida pela Organização para a Cooperação e Desenvolvimento Econômico - OCDE, que buscou incluir características da IA generativa para a delimitação da sua concepção de sistema de IA.

A inclusão de conteúdos e termos da ciência da computação, na primeira parte, se mostrou essencial para uma melhor compreensão das soluções de IA utilizadas no âmbito do Direito, pois termos como *machine learning* e algoritmo não fazem parte do vocabulário dos juristas. Além disso, a inserção de um tópico sobre a ascensão da IA generativa revelou-se fundamental, tendo em vista as crescentes relações entre esse subcampo e o Direito, de modo que se constatou o seu potencial para afetar o mercado jurídico em tarefas mais complexas, propiciando que as ferramentas de IA excedam as funções de incremento em atividades repetitivas e maçantes.

A partir da análise das estatísticas disponibilizadas pelo Conselho Nacional de Justiça - CNJ, notou-se que nos últimos quatro anos o número de projetos de inteligência artificial em uso e em desenvolvimento no Poder Judiciário cresceu exponencialmente, sendo o maior aumento verificado entre os anos de 2021 e 2022. Observou-se que a maior parte das ferramentas são direcionadas aos servidores e usuários internos dos tribunais e as tarefas dos modelos comportam funções que vão da classificação de petições e agrupamento de processos por similaridade à transcrição de audiências, em uma lógica de apoio na qual não se pretende delegar funções decisórias à máquina.

Estabelecer regulações ao desenvolvimento e uso de modelos de inteligência artificial é um desafio que precisa ser enfrentado pelo Brasil. Esse debate tem se arrastado no Poder Legislativo nos últimos cinco anos e a discussão vem se ampliando com a tramitação de novos Projetos de Lei nas duas Casas Legislativas, os quais buscam acompanhar as constantes evoluções da IA. A aprovação do AI Act na União Europeia é um marco mundial para a regulação do tema e certamente servirá de referência para a aprovação de futuras leis nacionais. Enquanto não há aprovação de um marco legal, outros atos normativos se destacam pela

tentativa de estabelecer limites e parâmetros, tendo sido enfatizado no presente trabalho a Resolução CNJ n. 332/2020, que dispõe de maneira específica sobre a produção e o uso de IA no Poder Judiciário.

No capítulo destinado à ética e transparência da IA, devido à grande quantidade de trabalhos sobre o tema, priorizou-se os estudos de pesquisadores que fizeram revisões sistemáticas de literatura, a fim de se chegar a uma abordagem mais robusta para a questão. Detectou-se que os princípios da ética da IA são dinâmicos e a verificação destes depende do contexto e das atividades que serão executadas pelos sistemas. Assim, um *chatbot* não suscita as mesmas preocupações éticas que um sistema de inteligência artificial que será utilizado como apoio às decisões judiciais. Portanto, a depender do caso, verifica-se uma preponderância entre os princípios éticos.

Implementar os princípios éticos em um modelo de IA tende a ser problemático, pois o caminho entre os princípios e a prática pode ser lacunoso. Assim, investigou-se propostas para minimizar esse obstáculo, como a criação de avaliações de impacto, o esclarecimento de resultados, a criação de comitês de ética nas entidades, a elaboração de normas e a sensibilização de usuários e afetados para a existência desse arcabouço. Evidenciou-se em pesquisa realizada pelo CNJ (2023) que a principal medida adotada ou em vias de adoção em relação à transparência e ética no uso de IA nos tribunais é fomentar informações aos usuários sobre a existência de IA, o que enfatiza a importância da comunicação quanto ao uso de tais ferramentas.

Os projetos de Pesquisa e Desenvolvimento, enquanto fonte de consolidação de conhecimentos e posterior utilização dos recursos obtidos para a construção de novas aplicações com base em metodologia específica, são apropriados à criação de sistemas de inteligência artificial. Além disso, há uma relação estreita entre a Pesquisa e Desenvolvimento de ferramentas de IA no Direito e o cumprimento de referenciais éticos, que se traduz em uma maior confiança nos modelos. Identificou-se também que os investimentos em P&D feitos pelo Brasil precisam se expandir para que o país avance no quesito inovação, considerando-se a comparação com o cenário internacional.

As últimas seções do trabalho tiveram como base o Projeto de Pesquisa e Desenvolvimento Sabiá. Dirigido ao ramo trabalhista da justiça especializada, o Projeto tem como objetivo desenvolver soluções de inteligência artificial para o Tribunal Superior do Trabalho visando uma melhor gestão do acervo processual, otimização dos recursos humanos

e aprimoramento da prestação jurisdicional, por meio de parceria com a Universidade de Brasília. De modo semelhante a outros tribunais, o TST enfrenta os problemas da crescente litigiosidade e de um alto percentual de casos pendentes de julgamento, situação na qual o uso de ferramentas de IA pode ser estratégico.

Mesmo que ainda não tenha sido implementado pelo TST, é possível examinar aspectos éticos do Projeto Sabiá. Para isso, foram obtidas as respostas fornecidas pelo Tribunal a perguntas feitas no âmbito do Painel da Pesquisa sobre Inteligência Artificial 2023, divulgado pelo CNJ. Foi feita a opção pela aferição dessas informações em razão de estarem disponíveis ao público externo e pela facilidade de acessá-las. A partir disso, relacionou-se esse conteúdo a determinados princípios éticos da IA, selecionados a partir de uma análise preliminar feita no capítulo anterior. Concluiu-se que o conjunto de explicações fornecidas são necessárias, mas não suficientes ao adequado cumprimento de referenciais éticos, sendo que ajustes podem ser promovidos a fim de aperfeiçoar esse quadro.

O problema de pesquisa da presente dissertação, que se refere à maneira pela qual o Projeto Sabiá relaciona-se aos parâmetros éticos necessários à Pesquisa e Desenvolvimento de um sistema de inteligência artificial, foi respondido na última parte. A reflexão empírica sobre a ética da IA é, de certa forma, um modo de trazer o jurisdicionado para o centro do debate, uma vez que direitos fundamentais estão em evidência e a promoção de um ambiente de transparência e robustez contribui para uma maior confiança de que os litígios serão resolvidos de forma justa e efetiva.

REFERÊNCIAS

ACIOLY, Luiz Henrique de Menezes; MENDES, Isabelle Brito Bezerra; MONTEIRO NETO, João Araújo. As Avaliações de Impacto como Instrumento de Inteligibilidade Algorítmica e Garantia de Direitos Fundamentais na Regulação de Inteligência Artificial. *Revista Jurídica Diké*, Uesc, v. 22, n 24, p. 225-251, 2023.

ADADI, Amina; BARRADA, Mohammed. Peekin Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). 2018. Disponível em: https://www.researchgate.net/publication/327709435_Peeking_Inside_the_Black-Box_A_Survey_on_Explainable_Artificial_Intelligence_XAI. Acesso em: 22 jun. 2024.

ANDRADE, Otávio Morato de; NUNES, Dierle. O potencial da inteligência artificial para a otimização do sistema de dimensionamento de conflitos. *Revista da UFMG*. Belo Horizonte, MG, 2023.

ANGWIN, Julia; LARSON, Jeff; MATTU, Surya; KIRCHNER, Lauren. *Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*. ProPublica, Maio 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 21 fev. 2024.

ARRIETA, Alejandro Barredo; DÍAZ-RODRÍGUEZ, Natalia; DEL SER, Javier; BENNETOT, Adrien; TABIK, Siham; BARBADO, Alberto; GARCIA, Salvador; GIL-LOPEZ, Sergio; MOLINA, Daniel; BENJAMINS, Richard; CHATILA, Raja; HERRERA, Francisco. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion - Elsevier*, v. 58, p. 82-115, 2020.

ARTASANCHEZ, Alberto; JOSHI, Prateek. *Artificial Intelligence with Python: your complete guide to building inteligente apps using Python 3.x and TensorFlow 2*. Packt, Birmingham-Mumbai, 2nd edition, 2020.

AURÉLIEN, Géron. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow*. Editora Alta Books, 1ª ed., 2021.

BERGMANN, Dave. *O que é RLHF?*. IBM, 2023. Disponível em: https://www.ibm.com/br-pt/topics/rlhf#:~:text=A%20aprendizagem%20por%20refor%C3%A7o%20a,meio%20da%20aprendizagem%20por%20refor%C3%A7o. Acesso em: 10 mar. 2024.

BERTOLINI, Andrea. *Artificial Intelligence and Civil Liability - Legal Affairs*. Policy Department for Citizen's Rights and Constitutional Affairs. Brussels, 2020. Disponível em: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf. Acesso em: 17 jan. 2024.

BONAT, Debora; PEIXOTO, Fabiano Hartmann. Racionalidade no Direito: Inteligência Artificial e Precedentes. Coleção Direito, Racionalidade e Inteligência Artificial – volume 3. Curitiba: Alteridade, edição Kindle, 2020.

BONAT, Debora; HARTMANN PEIXOTO, Fabiano. Inteligência Artificial e processo judicial: otimização comportamental e relação de apoio. *Revista Humanidades e Inovação*, Universidade Estadual do Tocantins (Unitins), v. 8, n. 47, 2021.

BONAT, Debora; HARTMANN PEIXOTO, Fabiano. GPTs e Direito: impactos prováveis das IAs generativas nas atividades jurídicas brasileiras. *Sequência (Florianópolis)*, v. 44, n. 93, 2023.

BRAGANÇA, Fernanda; LOSS, Juliana; BRAGA, Renata. Tecnologia na Justiça. In: SALOMÃO, Luis Felipe (Coord.) *Inteligência Artificial – Tecnologia Aplicada à Gestão dos Conflitos no âmbito do Poder Judiciário Brasileiro*. 2ª edição. Rio de Janeiro: Centro de Inovação, Administração e Pesquisa do Judiciário da Fundação Getúlio Vargas, 2022.

BRASIL. *Projeto de Lei n. 21, de 2020*. Atividade Legislativa - Tramitação Câmara dos Deputados. 2021. Disponível em: https://www.camara.leg.br/propostas-legislativas/2236340. Acesso em: 12 mar. 2024.

BRASIL. *Projeto de Lei n. 21, de 2020*. Atividade Legislativa - Tramitação Senado Federal. 2024. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/151547. Acesso em: 13 mar. 2024.

BRASIL. *Portaria MCTI n. 4.617, de 6 de abril de 2021*- Institui a Estratégia Brasileira de Inteligência Artificial e seus eixos temáticos. 2021. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-portaria_mcti_4-617_2021.pdf. Acesso em: 10 jan. 2024.

BRASIL. *Projeto de Lei n. 2.338, de 2023.* Atividade Legislativa - Tramitação Senado Federal. 2023. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 20 mar. 2024.

BRASIL. *Projeto de Lei n. 145, de 2024*. Atividade Legislativa - Tramitação Senado Federal. 2024(b). Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/161946. Acesso em: 05 abr. 2024.

BRASIL. *Projeto de Lei n. 146, de 2024*. Atividade Legislativa - Tramitação Senado Federal. 2024 (c). Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/161947. Acesso em: 05 abr. 2024.

BRASIL. *Projeto de Lei n. 210, de 2024*. Atividade Legislativa - Tramitação Senado Federal. 2024 (d). Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/161980. Acesso em: 05 abr. 2024.

BRASIL. *Projeto de Lei n. 266, de 2024*. Atividade Legislativa - Tramitação Senado Federal. 2024 (e). Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/162045. Acesso em: 07 abr. 2024.

BRASIL. *Busca - Portal do Senado Federal*. 2024 (f). Disponível em: https://www6g.senado.leg.br/busca/?q=intelig%C3%AAncia+artificial&colecao=Projetos+e+

Mat%C3%A9rias+-+Proposi%C3%A7%C3%B5es&tipo-materia=PL+-+Projeto+de+Lei&p=3. Acesso em: 20 jun. 2024.

BRASIL. *Projetos de Lei e Outras Proposições*. Câmara dos Deputados. 2024 (g), Disponível em:https://www.camara.leg.br/buscaProposicoesWeb/resultadoPesquisa?numero=&ano=&autor=&inteiroTeor=intelig%C3%AAncia+artificial&emtramitacao=Sim&tipoproposicao=%5BPL+-+Projeto+de+Lei%5D&data=25/06/2024&page=false. Acesso em: 20 jun. 2024.

CAMPOS, André Gambier; DI BENEDETTO, Roberto. Mercado de trabalho jurídico no Brasil: Qual é a situação atual?. *Texto para Discussão*, IPEA, 2021.

CEPEJ (Comissão Europeia para a eficácia da Justiça). *Carta Europeia de Ética sobre o Uso da Inteligência Artificial em Sistemas Judiciais e seu ambiente*. 31ª reunião plenária, Estrasburgo, 2018. Disponível em: https://rm.coe.int/carta-etica-traduzida-para-portugues-revista/168093b7e0. Acesso em: 15 mar. 2024.

CHADHA, Anupama; KUMAR, Vaibhav; KASHYAP, Sonu; GUPTA, Mayank. Deepfake: An Overview. *In: Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*. Lecture Notes in Networks and Systems, vol. 203. Springer, Singapore, 2021.

CIHON, Peter. *Standards foi AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. University of Oxford. 2019. Disponível em: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf. Acesso em: 25 jun. 2024.

CNJ (Conselho Nacional de Justiça). *Resolução n. 185 de 18/12/2013*. 2013. Disponível em: https://atos.cnj.jus.br/atos/detalhar/1933. Acesso em: 05 mar. 2024.

CNJ (Conselho Nacional de Justiça). *Resolução n. 335 de 29/09/2020*. 2020. Disponível em: https://atos.cnj.jus.br/atos/detalhar/3496. Acesso em: 05 mar. 2024.

CNJ (Conselho Nacional de Justiça). *Resolução n. 332 de 21/08/2020*. 2020(b). Disponível em: https://atos.cnj.jus.br/atos/detalhar/3429. Acesso em: 05 mar. 2024.

CNJ (Conselho Nacional de Justiça). *Portaria n. 271 de 04/12/2020*. 2020(c). Disponível em: https://atos.cnj.jus.br/atos/detalhar/3613. Acesso em: 08 mar. 2024.

CNJ (Conselho Nacional de Justiça). *Portaria n. 36 de 14/02/2023*. 2023(b). Disponível em: https://atos.cnj.jus.br/atos/detalhar/4953. Acesso em: 10 mar. 2024.

CNJ (Conselho Nacional de Justiça). *Portaria n. 338 de 30/11/2023*. 2023(c). Disponível em: https://atos.cnj.jus.br/atos/detalhar/5368. Acesso em: 12 mar. 2024.

CNJ (Conselho Nacional de Justiça). *Ética e Transparência - Plataforma Sinapses*. 2021. Disponível em: https://www.cnj.jus.br/sistemas/plataforma-sinapses/etica-e-transparencia/#:~:text=332%2F2020%2C%20que%20trata%20sobre,de%20decis%C3%A3o%20nos%20%C3%B3rg%C3%A3os%20judiciais. Acesso em: 03 abr. 2024.

CNJ (Conselho Nacional de Justiça). Justiça 4.0. [s.d]. Disponível em: https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/. Acesso em 07 mar. 2024.

CNJ (Conselho Nacional de Justiça). *Painel de Projetos de IA no Poder Judiciário - 2022*. Disponível em: https://paineisanalytics.cnj.jus.br/single/?appid=9e4f18ac-e253-4893-8ca1-b81d8af59ff6&sheet=b8267e5a-1f1f-41a7-90ff-d7a2f4ed34ea&lang=pt-BR&theme=IA_PJ&opt=ctxmenu,currsel&select=language,BR. Acesso em: 06 abr. 2024.

CNJ (Conselho Nacional de Justiça). *Painel de Projetos de IA no Poder Judiciário - 2023*. Disponível em: https://paineisanalytics.cnj.jus.br/single/?appid=43bd4f8a-3c8f-49e7-931f-52b789b933c4&sheet=e4072450-982c-48ff-9e2d-361658b99233&theme=horizon&lang=pt-BR&opt=ctxmenu,currsel&select=Ramo%20da%20Justi%C3%A7a,&select=Tribunal,&select=Seu%20Tribunal/%20Conselho%20possui%20Projeto%20de%20IA?,.Acesso em: 15 maio 2024.

CNJ (Conselho Nacional de Justiça). *Justiça em Números 2024 (ano-base 2023)*. 2024. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2024/05/justica-em-numeros-2024.pdf. Acesso em 05 jun. 2024.

CORREA, Cristina Mendes Bertoncini; GONCALVES, Jéssica. Obstáculos históricos e simbólicos à transformação da cultura de tratamento dos conflitos da sentença em solução consensual no sistema jurídico brasileiro. In: José Sérgio da Silva Cristóvam; Norma Sueli Padilha; Ubaldo Cesar Balthazar. (Org.). *Direito, Estado e Sociedade - Homenagem aos 50 anos do PPGD/UFSC*. 1ed.São Paulo: Matrioska Editora, v. I, p. 293-308,2022.

COZMAN, Fabio Gagliardi; KAUFMAN, Dora. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. *Revista USP*, n. 135, p. 195-210, 2022.

CSJT (Conselho Superior da Justiça do Trabalho). Bem-Te-Vi. [s.d]. Disponível em: https://www.csjt.jus.br/web/csjt/justica-4-0/bem-ti-vi. Acesso em: 25 jul. 2024.

DEAN, Jeffrey. A Golden Decade of Deep Learning: Computing Systems & Applications. *Daedalus, the Journal of the American Academy of Arts & Sciences*, v. 151 (2), Spring 2022.

EBIA. *Estratégia Brasileira para Inteligência Artificial*. Documento de referência, MCTI, 2021. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-diagramacao_4-979_2021.pdf. Acesso em: 10 jan. 2024.

E-DIGITAL. *Estratégia Brasileira para a Transformação Digital*. Brasília, 2018. Disponível em: https://www.gov.br/mcti/pt-br/centrais-de-conteudo/comunicados-mcti/estrategia-digital brasileira/estrategiadigital.pdf. Acesso em: 15 dez. 2023.

EUROPEAN PARLIAMENT. *Artificial Intelligence Act*. Texts Adopeted. 2024. Disponível em: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf. Acesso em: 10 maio de 2024.

EUROPEAN PARLIAMENT. *Artificial Intelligence Act: MEPs adopt landmark law*. 2024 (b). Disponível em: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law . Acesso em: 10 maio de 2024.

EWALD, John. *Introduction to large language models*. YouTube, 2023, Disponível em: https://www.youtube.com/watch?v=zizonToFXDs. Acesso em: 10 mar. 2024.

FOSTER, David. *Generative Deep Learning*. O'Reilly, 2nd edition, 2023.

GASPAR, Walter Britto; MENDONÇA, Yasmin Curzi. *Inteligência Artificial no Brasil ainda precisa de uma estratégia*. Portal FGV, 2021. Disponível em: https://portal.fgv.br/artigos/inteligencia-artificial-brasil-ainda-precisa-estrategia. Acesso em: 12 jan. 2024.

GRYNBAUM, Michael M.; MAC, Ryan. *The Times Sues OpenAI and Microsoft over AI Use of Copyrighted Work*. The New York Times, 2023. Disponível em: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html. Acesso em: 10 mar. 2024.

HAO, Karen. *This is how AI bias really happens - and why it's so hard to fix*. MIT Technology Review, 2019. Disponível em: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/Acesso em: 25 jan. 2024.

HAENLEIN, Michael; KAPLAN, Andreas. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *BerkleyHaas*. Vol. 61(4) 5-14, 2019.

HARTMANN PEIXOTO, Fabiano; SILVA, Roberta Zumblick Martins da. *Inteligência Artificial e Direito*. 1 ed., Coleção Direito, Racionalidade e Inteligência Artificial. Curitiba: Alteridade Editora, 2019.

HARTMANN PEIXOTO, Fabiano. *Direito e Inteligência Artificial – referenciais básicos*. Coleção Inteligência Artificial e Jurisdição, DOI 10.29327/521174, volume 2, 2020(a).

HARTMANN PEIXOTO, Fabiano. Projeto Victor: relato do desenvolvimento da inteligência artificial na repercussão geral do Supremo Tribunal Federal. *Revista Brasileira de Inteligência Artificial e Direito - RBIAD*, n. 1, ed. 1, 2020(b).

HARTMANN PEIXOTO, Fabiano. *Inteligência Artificial e Direito: Convergência Ética e Estratégica*. Coleção Direito, Racionalidade e Inteligência Artificial – volume 5. Curitiba: Alteridade, edição Kindle, 2020(c).

HARTMANN PEIXOTO, Fabiano. *Nota de Colaboração – Substitutivo ao PROJETO DE LEI Nº 21/2020*. Grupo de Pesquisa DR.IA. UnB - Direito e Inteligência Artificial. 2021. Disponível em: https://drive.google.com/file/d/16m1FVCQYHnv4E2ZueCoXEGUaJ8zzAyNX/view. Acesso em: 02 abr. 2024.

HUANG, Changwu; ZHANG, Zeqi; MAO, Bifei; YAO, Xin. An Overview of Artificial Intelligence Ethics. *IEE Transactions on Artificial Intelligence*, vol. 4, n. 4, 2023.

IBM. *What is a neural network?*. [s.d]. Disponível em: https://www.ibm.com/topics/neural-networks. Acesso em: 05 fev. 2024.

JOBIN, Anna; IENCA, Marcello; VAYENA, Effy. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. Vol. 1, n. 9, p. 389-399, 2019.

JUSTEN FILHO, Marçal. *Curso de Direito Administrativo*. Rio de Janeiro: Forense, Grupo Gen, 14ª ed, 2023.

KAHNEMAN, Daniel; SIBONY, Olivier; SUNSTEIN, Cass R. *Ruído: Uma falha no julgamento humano*. Rio de Janeiro: Objetiva, 2022.

KSIĘŻAK, Paweł; WOJTCZAK, Sylwia. *Toward a Conceptual Network for the Private Law of Artificial Intelligence.* Law, Governance and Technology. Series 51, Springer, 2023.

LAGE, Fernanda de Carvalho. *Manual de Inteligência Artificial no Direito Brasileiro*. Editora JusPodivm, 2021.

LAGE, Fernanda de Carvalho; HARTMANN PEIXOTO, Fabiano. Inteligência Artificial e Direito: desafios para a regulação do uso da inteligência artificial. Capítulo XII. In: HARTMANN PEIXOTO (Coord.). *Inteligência Artificial: Estudos de Inteligência Artificial*. Curitiba: Alteridade, 2021.

LEMOS, Aline Dayany; TORRES, Camila Ribeiro Rocha; ROCHA, Ana Carolina Pereira Rocha. Similaridade por assunto nos Despachos de Admissibilidade de processos trabalhistas. *In: II Congresso Nacional de Pesquisa Judiciária, Ciência de Dados e Estatística na Justiça do Trabalho* - GT 2 "Análise de Dados: técnicas e ferramentas voltadas à boa governança de dados". 2024. Disponível em:https://tst.jus.br/documents/31882359/0/Rela%C3%A7%C3%A3o+de+Grupos+Tem%C3%A1ticos+GT+2.pdf/96d7455a-19b2-9c81-ef4c-4fada63b3cff?t=1722463649500. Acesso em: 24 ago. 2024.

LITTMAN, Michael L.; AJUNWA, Ifeoma; BERGER, Guy; BOUTILIER, Craig; CURRIE, Morgan; DOSHI-VELEZ, Finale; HADFIELD, Gillian; HOROWITZ, Michael C.; ISBELL, Charles; KITANO, Hiroaki; LEVY, Karen; LYONS Terah, MITCHELL Melanie; SHAH, Julie; SLOMAN, Steven; VALLOR, Shannon; WALSH Toby. *"Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report".* Stanford University, Stanford, CA, Set. 2021. Disponível em: http://ai100.stanford.edu/2021-report. Acesso em: 15 dez. 2023.

MAINI, Vishal; SABRI, Samer. *Machine Learning for Humans*. 2017. Disponível em: https://everythingcomputerscience.com/books/Machine%20Learning%20for%20Humans.pdf. Acesso em: 05 fev. 2024.

MARCONDES, Danilo. *Textos básicos de ética: de Platão a Foucault*. Editora Zahar, 4ª edição, 2007.

MARTINS, Maria do Carmo. *Ada Lovelace: a primeira programadora da história*. 2016. Disponível em: https://repositorio.uac.pt/handle/10400.3/4025. Acesso em: 10 dez. 2023.

MOOR, James. The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine*, v. 27, n.4, 2006.

MORLEY, Jessica; FLORIDI, Luciano; KINSEY, Libby; ELHALAL, Anat. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, v. 26, n. 4, p. 2141-2168, 2020.

NONATO, Livia. *Pesquisa e Desenvolvimento (P&D): o que é e como fazer*. Aevo. 2023. Disponível em: https://blog.aevo.com.br/pesquisa-e-desenvolvimento/#o-que-e-pesquisa-e-desenvolvimento. Acesso em: 03 jun. 2024.

NUNES, Dierle. Virada tecnológica no direito processual: fusão de conhecimentos para geração de uma nova justiça centrada no ser humano. *Revista de Processo*. vol. 344. ano 48. p. 403-429. São Paulo: Ed. RT, 2023.

NUNES, Dierle; MARQUES, Ana Luiza Pinto Coelho. Algoritmo: O risco da decisão das máquinas. *Revista Bonijuris*, vol. 31, n.4 - #659, pp. 44-58, ago/set 2019.

OCDE (Organização para a Cooperação e Desenvolvimento Econômico). OECD AI Principles overview. 2024. Disponível em: https://oecd.ai/en/ai-principles. Acesso em: 10 jun. 2024.

OCDE (Organização para a Cooperação e Desenvolvimento Econômico). *Research and development (R&D)*. OECD iLibrary, 2024 (b). Disponível em: https://www.oecd-ilibrary.org/industry-and-services/research-and-development-r-d/indicator-group/english_09614029-en. Acesso em: 29 jun. 2024.

OLIVEIRA, Cristina Godoy Bernardo de. Desafios da regulação do digital e da inteligência artificial no Brasil. *Revista USP*, n. 135, p. 137-162, 2022.

OPENAI. *Harvey*. 2024. Disponível em: https://openai.com/index/harvey/. Acesso em: 02 maio 2024.

PINHEIRO, Andréia Azevedo; SIANI, Antônio Carlos; GUILHERMINO, Jislaine de Fátima; HENRIQUES, Maria das Graças Muller de Oliveira; QUENTAL, Cristiane Machado; PIZARRO, Ana Paula Brum. Metodologia para gerenciar projetos de pesquisa e desenvolvimento com foco em produtos: uma proposta. *Revista de Administração Pública*, v. 40, p. 457-478, 2006.

REEVE, Octavia. *Celebrating Ada Lovelace Day: what Ada means to us*. Ada Lovelace Institute, 2019. Disponível em: https://www.adalovelaceinstitute.org/blog/celebrating-ada-lovelace-day/. Acesso em: 25 jan. 2024.

ROYCHOWDHURY, Sohini. *Journey of Hallucination-minimized Generative AI Solutions for Financial Decision Makers*. ACM International Conference on Web Search and Data Mining, 2023. Disponível em: https://arxiv.org/pdf/2311.10961.pdf . Acesso em: 15 mar. 2024.

RUSSELL, Stuart J.; NORVIG, Peter. *Inteligência Artificial.* Terceira Edição, Rio de Janeiro: Grupo Gen, 2013.

RUSSELL, Stuart J.; NORVIG, Peter. *Inteligência Artificial - Uma Abordagem Moderna.* Quarta Edição, Rio de Janeiro: Grupo Gen, 2022.

RUSSEELL, Stuart. If We Succed. *Daedalus, the Journal of the American Academy of Arts & Sciences*, v. 151 (2), Spring 2022.

RUSSELL, Stuart; PERSET, Karine; GROBELNIK, Marko. *Updates to the OECD's definition of an AI system explained*. 2023. Disponível em: https://oecd.ai/en/wonk/ai-system-definition-update. Acesso em: 02 fev. 2024.

RUSTER, Lorenn P.; OLIVA-ALTAMIRANO, Paola; DANIELL, Katherine A. Centring dignity in algorithm development: testing a Dignity Lens. In: *Proceedings of the 34th Australian Conference on Human-Computer Interaction*. 2022.

RYAN, Mark; STAHL, Bernd Carsten. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society.* Vol. 19, n. 1, 2021.

SALOMÃO, Luis Felipe (Coord.). *Inteligência Artificial – Tecnologia Aplicada à Gestão dos Conflitos no âmbito do Poder Judiciário Brasileiro*. Rio de Janeiro: Centro de Inovação, Administração e Pesquisa do Judiciário da Fundação Getúlio Vargas, 2020.

SALOMÃO, Luis Felipe (Coord.). *Inteligência Artificial – Tecnologia Aplicada à Gestão dos Conflitos no âmbito do Poder Judiciário Brasileiro*. 2ª edição, Rio de Janeiro: Centro de Inovação, Administração e Pesquisa do Judiciário da Fundação Getúlio Vargas, 2022.

SALOMÃO, Luis Felipe (Coord.). *Nota técnica - substituto ao Projeto de Lei 21/2020*. 2ª edição. Rio de Janeiro: Centro de Inovação, Administração e Pesquisa do Judiciário da Fundação Getúlio Vargas, 2022 (b).

SALOMÃO, Luis Felipe; TAUK, Caroline Somesom (Coord.). *Inteligência Artificial – Tecnologia Aplicada à Gestão dos Conflitos no âmbito do Poder Judiciário Brasileiro*. 3ª edição, Rio de Janeiro: Centro de Inovação, Administração e Pesquisa do Judiciário da Fundação Getúlio Vargas, 2023.

SALOMÃO, Luis Felipe; TAUK, Caroline Somesom. Inteligência Artificial no Judiciário Brasileiro: Estudo Empírico sobre Algoritmos e Discriminação. *Revista Jurídica Diké*, Uesc, v. 22, n 23, p. 02-32, 2023 (b).

SCHIFF, Daniel; RAKOVA, Bogdana; AYESH, Aladdin; FANTI, Anat; LENNON, Michael. Explaining the principles to practices gap in AI. *IEEE Technology and Society Magazine*, v. 40, n. 2, p. 81-94, 2021.

SHONO, Gessica. Software customizado vs. Software de prateleira: entenda as diferenças. Blog Ateliware, 2024. Disponível em: https://blog.ateliware.com/software-customizado-vs-software-de-prateleira/. Acesso em: 02 jul. 2024.

SICHMAN, Jaime Simão. Inteligência Artificial e sociedade: avanços e riscos. *USP-Estudos Avançados* 35(101), pp. 37-49, 2021. Disponível em: Acesso em: 03 jan. 2024.

SIQUEIRA, Dirceu Pereira; MORAIS, Fausto Santos; SANTOS, Marcel Ferreira dos. Inteligência artificial e jurisdição: dever analítico de fundamentação e os limites da substituição dos humanos por algoritmos no campo da tomada de decisão judicial. *Seqüência: estudos jurídicos e políticos*, v. 43, n. 91, 2022.

SIQUEIRA, Eduardo Camargo de; SILVA, Nilton Correia da; CORREA, Eduardo Ramos. Sabiá - Análise de Similaridade em Recursos Trabalhistas: Uma Abordagem de Aprendizado de Máquina e Ciência de Dados. *In: II Congresso Nacional de Pesquisa Judiciária, Ciência de Dados e Estatística na Justiça do Trabalho* - GT 2 "Análise de Dados: técnicas e ferramentas voltadas à boa governança de dados". 2024. Disponível em:https://tst.jus.br/documents/31882359/0/Rela%C3%A7%C3%A3o+de+Grupos+Tem%C3%A1ticos+GT+2.pdf/96d7455a-19b2-9c81-ef4c-4fada63b3cff?t=1722463649500. Acesso em: 24 ago. 2024.

STOKEL-WALKER, Chris. *Generative AI is coming for the lawyers*. Wired, 2023. Disponível em: https://www.wired.com/story/chatgpt-generative-ai-is-coming-for-the-lawyers/. Acesso em: 02 maio 2024.

STRIPLING, Gwendolyn; ABEL, Michael. *Low-Code AI: A Practical Project-Driven Introduction to Machine Learning*. O'Reilly, 1st edition, 2023.

STRIPLING, Gwendolyn. *Introduction to Genarative AI*. YouTube, 2023. Disponível em: https://www.youtube.com/watch?v=G2fqAlgmoPo. Acesso em: 10 mar. 2024.

SUSSKIND, Richard. *Online Courts and the future of justice*. Oxford University press, 2019.

SUTTON, Richard S.; BARTO, Andrew G. *Reinforcement Learning: An Introduction*. The MIT Press. Cambridge, Massachusetts. London, England. 2015. Disponível em: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf. Acesso em: 13 jan. 2023.

TSE (Tribunal Superior Eleitoral). *Resolução n. 23.732, de 27 de fevereiro de 2024*. 2024. Disponível em: https://www.tse.jus.br/legislacao/compilada/res/2024/resolucao-no-23-732-de-27-de-fevereiro-de-2024. Acesso em: 20 abr. 2024.

TSE (Tribunal Superior Eleitoral). *TSE proíbe uso de inteligência artificial para criar e propagar conteúdos falsos nas eleições*. 2024 (b). Disponível em: https://www.tse.jus.br/comunicacao/noticias/2024/Fevereiro/tse-proibe-uso-de-inteligencia-artificial-para-criar-e-propagar-conteudos-falsos-nas-eleicoes. Acesso em: 20 abr. 2024.

TST (Tribunal Superior do Trabalho). *Relatório Geral da Justiça do Trabalho 2023*. Coordenadoria de Estatística do TST. Brasília, DF. 2024. Disponível em: https://www.tst.jus.br/documents/18640430/31950226/RGJT2022.pdf/fa638cf6-969b-6508-09d8-625ffba9cd93?t=1689185086782. Acesso em: 03 jul. 2024.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan; KAISER, Lukasz; POLOSUKHIN, Illia. *Attention is all you need*. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

VICENT, James. *What a machine learning tool that turns Obama white can (and can't) tell us about AI bias*. The Verge, 2020. Disponível em: https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias. Acesso em: 30 jun. 2024.

WAELEN, Rosalie. Why AI ethics is a critical theory. *Philosophy & Technology*, v. 35, n. 1, 2022.

WIPO. World Intellectual Property Organization. *Global Innovation Index 2023: Innovation in the face of uncertainty*. Geneva, Switzerland. 2023. Disponível em: https://www.wipo.int/edocs/pubdocs/en/wipo-pub-2000-2023-en-global-innovation-index-2023-16th-edition.pdf. Acesso em: 03 jul. 2024.

XIANG, Alice. Mirror, Mirror, on the Wall, Who's the Fairest of Them All?. *Daedalus, the Journal of the American Academy of Arts & Sciences,* v. 153 (1), Winter 2024.

ZHOU, Jianlong; CHEN, Fang. *AI Ethics: From Principles to Practice*. Data Science Institute, University of Technology Sydney, Sydney, Australia. 2022. Disponível em: https://opus.lib.uts.edu.au/bitstream/10453/164551/2/ethics_position_AISociety2022_final_version.pdf . Acesso em: 10 jun. 2024.