Revista IBERC
v. 3, n. 3, p. 1-28, set./dez. 2020
www.responsabilidadecivil.org/revista-iberc
DOI: https://doi.org/10.37963/iberc.v3i2.133

DIRETRIZES ÉTICAS PARA A INTELIGÊNCIA ARTIFICIAL CONFIÁVEL NA UNIÃO EUROPEIA E A REGULAÇÃO JURÍDICA NO BRASIL

ETHICS GUIDELINES FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE IN EUROPEAN UNION AND LEGAL REGULATION IN BRAZIL

Gabriela Buarque Pereira Silva [i]
Marcos Ehrhardt Júnior [ii]

RESUMO: Para além da seleção de notícias no *feed* do *Facebook* ou das músicas indicadas no *Spotify,* a expansão da inteligência artificial no contexto social é essencialmente viral: está em constante variabilidade e em crescente propagação. Nesse ponto, quais devem ser os parâmetros éticos e jurídicos de regulação da inteligência artificial? Quais elementos tornam uma tecnologia minimamente confiável? Visando responder tais questionamentos, a União Europeia publicou, em 2019, documento com Diretrizes Éticas para a Inteligência Artificial Confiável. O presente texto visa descrever os elementos das Diretrizes, analisar seus conceitos e sua aplicabilidade no contexto do ordenamento jurídico brasileiro, e consagrar parâmetros mínimos de controle do desenvolvimento tecnológico.

Palavras-chave: Inteligência Artificial. Ética. Diretrizes. União Europeia. Regulação.

ABSTRACT: For beyond selection of news in Facebook feed or songs indication in Spotify, artificial intelligence expansion in social context is essentially viral: it's in constant variability and in increasing propagation. At this point, which should the ethical and legal parameters for artificial intelligence regulation? Which elements make a technology minimally reliable? In order to answer such questions, European Union published in 2019 a document with Ethics Guidelines for Trustworthy Artificial Intelligence. This text aims to describe Guideline's elements, analyzing their concepts and their applicability in Brazilian legal system, establishing minimum parameters of technological development control.

Keywords: Artificial intelligence. Ethics. Guidelines. European Union. Regulation.

[i] Assessora judiciária no Tribunal de Justiça de Alagoas. Mestranda em Direito Público pela Universidade Federal de Alagoas e membro do grupo de pesquisa Direito Privado e Contemporaneidade. Bacharel em Direito pela Faculdade de Direito da Universidade Federal de Alagoas. Atuou como estagiária na Justiça Federal em Alagoas, na Defensoria Pública da União e no escritório de advocacia Jairo e George Melo Advogados Associados e como advogada nos escritórios Jairo e George Melo Advogados Associados e Machado, Meyer, Sendacz e Opice Advogados. Foi monitora de Teoria Geral do Direito Civil e de Direito de Família, participante do projeto de extensão Observatório Jurídico: Combatendo Crimes Eleitorais e membro do Núcleo do Direito da Propriedade Intelectual, como participante do projeto de extensão Educação e Difusão da Propriedade Intelectual na Comunidade Escolar da Faculdade de Direito de Alagoas. ORCID https://orcid.org/0000-0002-9418-241X

[ii] Advogado. Doutor em Direito pela Universidade Federal de Pernambuco (UFPE) e Mestre pela Universidade Federal de Alagoas (UFAL). Professor de Direito Civil dos cursos de mestrado e graduação da Faculdade de Direito da Universidade Federal de Alagoas. Pesquisador Visitante do Instituto Max-Planck de Direito Privado Comparado e Internacional (Hamburgo/Alemanha). Líder do Grupo de Pesquisa Direito Privado e Contemporaneidade (UFAL). Editor da Revista Fórum de Direito Civil (RFDC). Diretor Regional Nordeste do Instituto Brasileiro de Direito Civil (IBDCIVIL). Membro do Instituto Brasileiro de Direito de Família (IBDFAM). ORCID https://orcid.org/0000-0003-1371-5921

1 INTRODUÇÃO

A pretensão cognoscente da disciplina jurídica de novos fenômenos sociais muitas vezes demanda do intérprete a revisitação de conceitos doutrinários e da epistemologia consolidada sobre determinada temática. O desafio se torna ainda mais complexo quando se trata se definir diretrizes éticas e axiológicas para o desenvolvimento de determinada atividade.

A ascensão da inteligência artificial no cotidiano – desde situações mais banais como *Spotify*[1]*, feed* de notícias de rede social, *Waze*[2]*,* até contextos mais sofisticados como transações no mercado financeiro e veículos autônomos – criou a necessidade da consolidação de parâmetros mínimos de verificação de confiabilidade desses sistemas.

O Grupo Europeu de Ética na Ciência e Novas Tecnologias, organização independente e multidisciplinar composta por especialistas designados pela Comissão Europeia, publicou as Diretrizes Éticas para a Inteligência Artificial Confiável, documento que visa estabelecer parâmetros mínimos para a aferição da confiabilidade do sistema tecnológico.

O presente trabalho visa, por meio de análise documental e metodologia dedutiva de revisão bibliográfica, descrever as diretrizes consagradas no texto, analisando os conceitos estipulados e a sua aplicabilidade no contexto do ordenamento jurídico brasileiro.

2 DA INTELIGÊNCIA ARTIFICIAL

Conceituar inteligência artificial não é tarefa fácil. Com efeito, não há um único conceito aceito de modo universal, existindo, no entanto, algumas características que singularizam o sistema como tal. Russel e Norvig, na obra *Artificial Intelligence: a modern approach,* listam as quatro maiores categorias onde se costuma conceituar a inteligência artificial, enquadrando-as em “sistemas que pensam como humanos”, “sistemas que agem como humanos”, “sistemas que pensam racionalmente” e “sistemas que agem racionalmente”.[3]

No que se refere aos aspectos que a singularizam e lhe atribuem o aspecto de racionalidade análoga à dos seres humanos, argumenta-se que

> O primeiro é a comunicação. Pode-se comunicar com uma entidade inteligente. Quanto mais fácil for se comunicar com uma entidade, mais inteligente a entidade parece. Pode-se comunicar com um cachorro, mas não sobre a Teoria da Relatividade de Einstein. O segundo é o conhecimento interno. Espera-se que uma entidade inteligente tenha algum conhecimento sobre si mesma. O terceiro é o conhecimento externo. Espera-se que uma entidade

[1] Spotify é um serviço de compartilhamento de músicas lançado em 2008, que sugere músicas ao usuário de acordo com suas preferências musicais.
[2] Waze é um serviço para dispositivos móveis baseado na navegação por satélite e que contém informações sobre rotas e mapas.
[3] Cf. NORVIG, Peter; RUSSELL, Stuart J. *Artificial Intelligence*: A Modern Approach. New Jersey: Prentice Hall, 1995.

> inteligente conheça o mundo exterior, para aprender sobre isso, e utilizar essa informação. A quarta é o comportamento orientado por objetivos. Espera-se que a entidade tome medidas para atingir seus objetivos. O quinto é a criatividade. Espera-se que uma entidade inteligente tenha algum grau de criatividade. Neste contexto, criatividade significa a capacidade de tomar uma ação alternativa quando a ação inicial falha. Uma mosca tenta sair de uma sala e as colisões contra a vidraça continuam a repetir o mesmo comportamento fútil. Quando um robô Al bate em uma janela, ele tenta sair usando a porta. A maioria das entidades Al possui esses cinco atributos por definição. (Tradução livre). [4-5]

Patrick Henry Winston, por sua vez, aduz que existem várias formas de definir a inteligência artificial, definindo-a como o estudo da computação que possibilita perceber, racionar e agir.[6] Ressalte-se, ainda, que muitas máquinas são conduzidas por interfaces de comandos, o que atrela sua atividade à vontade do emissor ou proprietário. Outras, no entanto, têm demonstrado grau de interatividade mais baixo, evidenciando condução mais autônoma em relação ao ser humano.

Desse modo, a condução da atividade da máquina difere entre os sistemas que possuem interatividade alta com o operador-usuário, usualmente se subordinando às suas emissões, e os sistemas que possuem interatividade baixa com o operador-usuário, usualmente demonstrando autossuficiência na condução das atividades.

Argumenta-se que existem três tipos de inteligência artificial.[7]A primeira seria a *Artificial Narrow Inteligence* (ANI), uma espécie de inteligência artificial que se especializa numa única área e possui um único objetivo definido, tais como as máquinas treinadas para jogar xadrez, por exemplo. A segunda seria a *Artificial General Intelligence* (AGI), inteligência que mimetiza a mente humana e tem várias habilidades, tais como planejar e resolver problemas, compreender ideias complexas e aprender rapidamente por meio da experiência. Por fim, a terceira seria a *Artificial Super Intelligence (*ASI), intelecto que seria mais inteligente que até mesmo o cérebro humano em diversas áreas, incluindo habilidades sociais.

Cabe ressaltar que a tendência contemporânea é que tais máquinas possuam cada vez mais autossuficiência, sendo a inteligência artificial uma demonstração da capacidade de reprodução cognitiva das máquinas em que o acúmulo de aprendizado visa simular a

---

[4] HALLEVY, Gabriel. The criminal liability of artificial intelligence entities- from Science fiction to legal social control. *Akron Intellectual Property Journal*, Akron, v. 4, p. 171-199, 2016. p. 175

[5] "The first is communication. One can communicate with an intelligent entity. The easier it is to communicate with an entity, the more intelligent the entity seems. One can communicate with a dog, but not about Einstein's Theory of Relativity. The second is internal knowledge. An intelligent entity is expected to have some knowledge about itself. The third is external knowledge. An intelligent entity is expected to know about the outside world, to learn about it, and utilize that information. The fourth is goal-driven behavior. An intelligent entity is expected to take action in order to achieve its goals. The fifth is creativity. An intelligent entity is expected to have some degree of creativity. In this context, creativity means the ability to take alternate action when the initial action fails. A fly tries to exit a room and bumps into a windowpane continues to repeat the same futile behavior. When an **Al** robot bumps into a window, it tries to exit using the door. Most **Al** entities possess these five attributes by definition".

[6] WINSTON, Patrick Henry. *Artificial Intelligence.* 3. ed. Boston: Addison-Wesley Publishing Company, 1993. p. 5.

[7] STRELKOVA, O. PASICHNYK, O. *Three types of artificial intelligence.* Disponível em: http://eztuir.ztu.edu .ua/jspui/bitstream/123456789/6479/1/142.pdf. Acesso em: 3 maio 2019.

experiência mental humana. Impende observar, contudo, que utilizar a experiência humana como parâmetro para a análise da IA pode acarretar dificuldades.

Nesse sentido, Jerry Kaplan[8] argumenta que

> há outro problema com o uso de recursos humanos como um critério para a inteligência artificial. Máquinas são capazes de realizar muitas tarefas que as pessoas não podem fazer, e muitas dessas performances certamente parecem demonstrações de inteligência. Um programa de segurança pode suspeitar de um ataque cibernético baseado em um padrão incomum de solicitações de acesso aos dados em um período de apenas quinhentos milissegundos; um sistema de alerta de tsunamis pode soar um alarme baseado em mudanças quase imperceptíveis nas alturas do oceano que refletem a geografia submarina complexa; Um programa de descoberta de drogas pode propor uma nova mistura, encontrando um padrão previamente despercebido de arranjos moleculares em compostos de tratamento de câncer bem-sucedidos. O comportamento exibido por sistemas como esses, que se tornará cada vez mais comum no futuro próximo, não se presta a comparação com as capacidades humanas[9].

Jerry Kaplan[10] acrescenta que a essência da inteligência artificial – na verdade, a essência da inteligência – é a capacidade de fazer generalizações apropriadas em tempo hábil, com base em dados limitados[11]. No contexto contemporâneo, a inteligência artificial assume espaço em diversos ramos e possui inúmeras funções, podendo ajudar especialistas a resolver difíceis problemas de análise, a desenvolver novas ferramentas, a aprender por meio de exemplos e representações, a trabalhar com estruturas semânticas e a criar novas oportunidades de mercado.[12] A inteligência artificial alastra-se de modo exponencial no cotidiano, desde atividades mais banais até atividades mais sofisticadas, sem que muitas vezes as pessoas se deem conta da utilização dessa tecnologia.

Desde quando acordamos com o *smartphone,* a música que escolhemos no *Spotify,* o trajeto que escolhemos no *Waze,* o filme que selecionamos na *Netflix*[13]*,* o *Uber*[14] que pedimos, até o cupom de desconto que é emitido em drogarias. No contexto contemporâneo, pouca coisa consegue escapar do alcance da tecnologia, o que vem nos tornando cada vez mais

---

[8] KAPLAN, Jerry. *Artificial Intelligence*: What everyone needs to know. Oxford: Oxford University Press, 2016. p. 4.

[9] "But there's another problem with using human capabilities as a yardstick for AI. Machines are able to perform lots of tasks that people can't do at all, and many such performances certainly feel like displays of intelligence. A security program may suspect a cyber attack based on an unusual pattern of data access requests in a span of just five hundred milliseconds; a tsunami warning system may sound an alarm based on barely perceptible changes in ocean heights that mirror complex undersea geography; a drug discovery program may propose a novel admixture by finding a previously unnoticed pattern of molecular arrangements in successful cancer treatment compounds. The behavior exhibited by systems like these, which will become ever more common in the near future, doesn't lend itself to comparison with human capabilities. Nonetheless, we are likely to regard such systems as artificially intelligent."

[10] KAPLAN, Jerry. *Artificial Intelligence*, cit., p. 5

[11] "The essence of AI – indeed the essence of intelligence – is the ability to make appropriate generalizations in a timely fashion based on limited data."

[12] WINSTON, Patrick Henry. *Artificial Intelligence*, cit., p. 10-14.

[13] Netflix é uma provedora de compartilhamento de filmes e séries, que sugere programas de acordo com as preferências do usuário.

[14] Uber é uma prestadora de serviços na área do transporte privado urbano, que atua por meio de um aplicativo de transporte que permite a busca por motoristas baseada na localização.

dependentes desses mecanismos, que essencialmente criam necessidades que até então nem sequer eram percebidas pelos seres humanos.

Um sistema de inteligência artificial não é somente capaz de armazenar e manipular dados, mas também de adquirir, representar e manipular conhecimento. Essa manipulação inclui a capacidade de deduzir novos conhecimentos a partir daqueles já existentes e utilizar métodos de representação para resolver questões complexas.[15]

Para Peter Norvig e Stuart Russell[16] a definição de um agente racional ideal se caracteriza quando, "para cada possível sequência de percepção, um agente racional ideal deve fazer qualquer ação que seja esperada para maximizar sua medida de desempenho, com base nas evidências fornecidas pela sequência perceptiva e em qualquer conhecimento embutido que o agente tenha". A inteligência artificial é, nesse aspecto, um mecanismo de acúmulo e representação de conhecimento, que se expande à medida que coleta mais dados.

Para isso, a inteligência artificial muitas vezes se utiliza de algoritmos, ferramenta que pode ser compreendida como uma sequência de etapas utilizada pela inteligência artificial para solucionar um problema ou realizar uma atividade. Os algoritmos podem atuar por meio de *machine learning*, que é, essencialmente, a atividade da máquina de aprender novos fatos por meio da análise dos dados e da experiência prévia, sem programação explícita para tanto, adaptando a aprendizagem a novas situações.[17]

Uma das inquietações oriundas das revoluções tecnológicas e industriais iniciadas no século XX, e que se avulta cada vez mais no contexto social, é a preocupação com a demarcação de limites na interação entre ser humano e inteligência artificial.

Em sede literária, por exemplo, Isaac Asimov enumerou as chamadas "Três Leis da Robótica" em sua obra "Eu, Robô": a) um robô não pode ferir um ser humano ou, por omissão, permitir que um ser humano sofra algum mal; 2) um robô deve obedecer às ordens que lhe sejam dadas por seres humanos, exceto nos casos em que tais ordens contrariem a Primeira Lei; c) um robô deve proteger a sua própria existência, desde que tal proteção não entre em conflito com a Primeira e a Segunda Leis.

A preocupação se renova cada dia quando se constata que a inteligência artificial assume cada vez mais espaço. Exsurge, destarte, um novo paradigma operacional cibernético cada vez mais presente, com máquinas tomando decisões e assumindo posturas típicas de indivíduos, onde funcionavam profissões ora obsoletas. Sistemas decidem como serão feitos os investimentos de um banco, carros são conduzidos de modo autônomo, negócios jurídicos são firmados por meio de *softwares* em contratos eletrônicos, microscópios da *Google Brain* são capazes de diagnosticar câncer[18], robôs são produzidos para colaborar no cotidiano de idosos

---

[15] CÂMARA, Marco Sérgio Andrade Leal Câmara. *Inteligência artificial*: representação de conhecimento. Disponível em: https://student.dei.uc.pt/~mcamara/artigos/inteligencia_artifici- al.pdf. Acesso em: 22 set. 2018. p. 1.

[16] NORVIG, Peter; RUSSELL, Stuart J. *Artificial Intelligence*, cit., p. 33.

[17] CERKA, Paulius; GRIGIENE, Jurgita; SIRBIKYTE, Gintare. Liability for damages caused by artificial intelligence. *Computer Law and Security Review*, Londres, v. 31, n. 3, p. 376-389, jun. 2015. p. 380.

[18] Disponível em: https://www.tecmundo.com.br/produto/129343-microscopio-google-realidade-aumentada-ia-detectar-cancer.htm. Acesso em: 20 set. 2018.

no Japão[19], além de mecanismos utilizados no cotidiano como *Spotify*, *Waze* e *Netflix.* São apenas amostras do potencial transformador da inteligência artificial no meio comunitário.

É inquestionável que o advento de novas descobertas científicas enseja a incerteza acerca de seus efeitos futuros, máxime ante o exponencial potencial que tais tecnologias costumam ostentar. Nesse panorama, visando direcionar o desenvolvimento de tal tecnologia, no dia 8 de abril de 2019 a Comissão Europeia divulgou diretrizes éticas para a inteligência artificial (IA) confiável, documento que se baseia no trabalho do Grupo Europeu de Ética na Ciência e Novas Tecnologias e outros esforços similares. A Comissão Europeia é instituição que, entre outras funções, propõe legislações e programas de ação no contexto europeu.

O Grupo Europeu de Ética na Ciência e Novas Tecnologias é uma organização independente e multidisciplinar, composta por especialistas designados pela Comissão Europeia, criada em 20 de novembro de 1991. Visa estudar os aspectos políticos e legislativos de cruzamento entre as dimensões éticas e sociais dos direitos humanos com o desenvolvimento tecnológico e científico.

O objetivo das Diretrizes em análise é promover uma inteligência artificial que seja confiável, característica que se desdobra em três componentes, que devem ser atendidos durante todo o ciclo de vida do sistema e necessariamente em conjunto: a) observância à legalidade; b) à ética; e c) à robustez, tanto do ponto de vista técnico como do ponto de vista social. É evidente que podem surgir tensões entre tais componentes, tendo-se a ponderação como técnica para alinhá-los.

As Diretrizes articulam, em seu primeiro capítulo, os direitos fundamentais e os princípios éticos cruciais no contexto de desenvolvimento da inteligência artificial. No segundo capítulo são elencados requisitos que os sistemas de inteligência artificial devem observar para que sejam considerados confiáveis. Por fim, no terceiro capítulo, as Diretrizes fornecem uma lista de avaliação que ajuda a operacionalizar os requisitos antes mencionados, sempre com a perspectiva de estimular as discussões atinentes ao tema.

3 DOS PRINCÍPIOS ÉTICOS DA INTELIGÊNCIA ARTIFICIAL

As Diretrizes determinam, de início, que a inteligência artificial deve respeitar a autonomia humana, a prevenção de danos, a justiça e a explicabilidade. Também deve observar a situação de grupos vulneráveis, como crianças e adolescentes, idosos, pessoas com deficiências ou outros marcados por assimetrias de poder e informação, tais como consumidores e trabalhadores.

Não obstante os inegáveis progressos proporcionados pelo desenvolvimento tecnológico, a inteligência artificial também comporta riscos e pode acarretar efeitos negativos,

[19] Disponível em: http://g1.globo.com/tecnologia/noticia/2011/10/robos-poderao-ajudar-populacao-de-idosos-no-japao-no-futuro.html Acesso em: 20 set. 2018.

inclusive impactos difíceis de mensurar ou identificar no presente momento. Faz-se imprescindível a adoção de medidas adequadas para mitigar tais riscos[20].

É fundamental que haja um conhecimento mínimo acerca das capacidades e limitações da inteligência artificial, com o objetivo de facilitar a rastreabilidade e a auditabilidade dos sistemas de IA, especialmente em situações críticas. Não se ignora que muitas vezes não há conhecimento exato acerca de como essas máquinas funcionam, fenômeno chamado de *black box* da inteligência artificial. Em comentário a tal fenômeno, Will Knight argumenta que "Nós podemos construir esses modelos, mas nós não sabemos como eles trabalham[21]".

A avaliação e a pesquisa nesse ramo, portanto, são essencialmente dinâmicas e visam continuamente a implementar requisitos e avaliar soluções, com o objetivo de que melhores resultados sejam alcançados para todas as partes envolvidas. As Diretrizes também ressaltam a necessidade de que os sistemas de inteligência artificial sejam centrados no ser humano, apoiados no compromisso de servir à humanidade e sua liberdade. Trata-se da perspectiva antropocêntrica que predomina no contexto pós-moderno, em que se alça a dignidade humana como epicentro do sistema normativo.

Uly de Carvalho[22] (2018, p. 97) argumenta que

> (...) cabe apenas reafirmar não oferecer consistência uma analogia com os seres humanos: a autonomia dos robôs, de natureza eminentemente tecnológica, reduz-se a uma capacidade de escolha, fomentada pelas potencialidades de uma variedade de combinações algorítmicas viabilizadas por um *software*. Não pressupõe em suas decisões um comportamento ético, realizável por uma ponderação valorativa e moral, e tão pouco contempla algum zelo com o outro, pois tal nota poderia destoar da própria eficiência que é a raiz da IA.

Autores como Vernor Vinge[23] suscitam questões de uma era de pós-humanidade baseada numa noção de singularidade tecnológica[24], na qual as máquinas adquirem um grande nível de autonomia. Já John Searle busca refutar a ideia de que uma máquina possa

---

[20] Disponível em: http://g1.globo.com/mundo/noticia/2015/07/robo-agarra-e-mata-trabalhador-dentro-de-fabrica-da-volkswagen.html. Acesso em: 26 set. 2018; Disponível em: https://gizmodo.uol.com.br/shopping-robos-acidente-crianca/ Acesso em: 26 set. 2018.

[21] "We can build these models but we don't know how they work". KNIGHT, Will. *The dark secret at the heart of AI.* Disponível em: https://www.technology review.com/s/604087/the-dark-secret-at-the-heart-of-ai/ Acesso em: 26 set. 2018.

[22] PORTO, Uly de Carvalho Rocha. *A responsabilidade civil extracontratual por danos causados por robôs autônomos.* 2018. Dissertação (Mestrado em Ciências Jurídico-Civilistas), Faculdade de Direito da Universidade de Coimbra, Coimbra, p. 97.

[23] VIGNE, Vernor. *What is the singularity?* Disponível em: https://www.frc.ri.cmu.edu/~hpm/book98/com.ch1/vinge.singularity.html Acesso em: 26 set. 2018.

[24] "Quando a inteligência maior que a humana impulsiona o progresso, esse progresso será muito mais rápido. De fato, parece não haver razão para que o progresso em si não envolva a criação de entidades ainda mais inteligentes – em uma escala de tempo ainda menor (…). Essa mudança será uma eliminação de todas as regras humanas, talvez em um piscar de olhos – uma fuga exponencial além de qualquer esperança de controle. Os desenvolvimentos que foram pensados para acontecer em 'um milhão de anos' provavelmente acontecerão no próximo século. É justo chamar este evento de uma singularidade ('a Singularidade' para os propósitos desta peça). É um ponto em que nossos modelos antigos devem ser descartados e uma nova realidade deve ser governada, um ponto que se tornará mais vasto que os assuntos humanos, até que a noção se torne um lugar-comum. No entanto, quando finalmente acontece, ainda pode ser uma grande surpresa e uma grande incerteza." (Tradução livre).

efetivamente pensar[25]. Verifica-se que, nesse contexto, muitas das tentativas de teorização culminam em abordagens mais ou menos polarizadas. O fato é que as Diretrizes norteiam a compreensão da inteligência artificial como um mecanismo para o bem-estar humano, servindo-lhe como instrumento de desenvolvimento.

Nesse diapasão, torna-se imprescindível que todas as diretrizes éticas e as obrigações legais aplicáveis aos processos de desenvolvimento, implantação e uso de inteligência artificial sejam devidamente observadas, além de ser o sistema de IA suficientemente robusto. A robustez se refere tanto à perspectiva técnica, que se consubstancia na adequação do sistema e no domínio de aplicação ou fase do ciclo de vida, como à perspectiva social, que considera o contexto e o ambiente em que o sistema opera.

No que tange aos direitos fundamentais a serem observados, impende ressaltar que os tratados da União Europeia e a Carta da União Europeia prescrevem uma série de direitos fundamentais que os Estados-membros e as instituições da União Europeia são obrigados a respeitar, no tocante à dignidade, liberdade, igualdade e solidariedade. O alicerce de tais direitos reside precisamente na dignidade humana, que atribui ao ser humano uma posição única e inalienável, central na vida social, política e econômica.

No caso do Brasil, verifica-se que é imprescindível que não haja violação às disposições do art. 5º da Constituição Federal, além daquelas decorrentes do regime e dos princípios adotados pela Carta. Não se ignora que podem exsurgir tensões em tal processo, máxime em se tratando de tipicidade aberta, o que demanda do intérprete um esforço argumentativo que compatibilize os interesses em questão.

Nesse contexto, elenca-se inicialmente a necessidade de respeito à dignidade da pessoa humana, enquanto sujeito que possui valor intrínseco e que não pode ser reprimido por tecnologias de inteligência artificial. Sob a perspectiva kantiana, surge a preocupação em tratar o ser humano como finalidade em si mesmo, rejeitando concepções que o reificam ou o tratam como instrumento a ser manipulado ou classificado. É imprescindível que os sistemas de inteligência artificial respeitem a integridade física e intelectual dos seres humanos.

É fundamental que seja respeitada a liberdade do indivíduo. O direito à liberdade se consubstancia na "prerrogativa fundamental que investe o ser humano de um poder de autodeterminação ou de determinar-se conforme a sua própria consciência[26]". Caracteriza-se pelo poder de atuação individual em busca de seus próprios objetivos, substancializado na liberdade de ação, de locomoção, de opinião, de expressão, de informação, de crença, de

[25] John Searle se utiliza, em 1980, de teste chamado de "O argumento do quarto chinês". Deduz que o robô possui limitações que o restringem no campo da sintaxe. O "argumento do quarto chinês" se refere à hipótese em que um indivíduo, falante apenas do idioma português, encontra-se fechado num quarto com uma caixa, símbolos em chinês e um livro com regras. Explicita-se que símbolos devem ser enviados quando outros são remetidos. Supõe-se que são enviadas sucessivas perguntas em chinês, de modo que o indivíduo sempre recorre ao material disponível, enviando respostas corretas em chinês, sem, contudo, compreender aquilo a que se refere. Por analogia, Searle argumenta que tal funcionamento se assemelha à computação, porquanto a máquina não possui capacidades cognitivas efetivas, limitando-se a gerenciar símbolos.

[26] CUNHA JÚNIOR, Dirley da. *Curso de Direito Constitucional.* 3. ed. rev. ampl. e atual., Salvador: Juspodivm, 2009. p. 664.

associação e opção profissional. Ainda no que tange à liberdade, esta deve ser compreendida não somente como a ausência de coerção sobre o indivíduo, mas também sob a perspectiva da ausência de vigilância injustificada e de manipulações indiretas.

Demanda-se, nesse teor, que o indivíduo esteja consciente de que está lidando com um mecanismo de inteligência artificial, dos dados que fornece para a tecnologia e do modo como tais dados serão utilizados. É nesse sentido que a Lei nº 13.709/18[27] (Lei Geral de Proteção de Dados) dispõe, em seu art. 9º, que o titular tem direito ao acesso facilitado às informações sobre o tratamento de seus dados, que deverão ser disponibilizadas de forma clara, adequada e ostensiva acerca de, entre outras características previstas em regulamentação para o atendimento do princípio do livre acesso, finalidade específica, forma e duração do tratamento, identificação e informações de contato do controlador, informações acerca do uso compartilhado de dados pelo controlador e finalidade, responsabilidades dos agentes que realizarão o tratamento e direitos do titular.

Trata-se de consagrar a autodeterminação informativa, de modo que o pleno exercício dos direitos de liberdade do indivíduo depende também do controle que possui acerca da circulação de seus dados, especialmente considerando que a manipulação dos dados pessoais pode acarretar uma representação que simboliza o indivíduo perante o meio social.

Ainda na perspectiva do direito à liberdade, é fundamental que o indivíduo detenha a capacidade de influenciar ou interromper o uso de seus dados pela inteligência artificial, não podendo haver utilização indevida ou que se desvirtue das finalidades propostas pelo usuário. Essa preocupação pode ser especialmente relevante na hipótese da comercialização de robôs sexuais, por exemplo. Isso porque muitas vezes tais robôs não são apenas máquinas estáticas, senão mecanismos que utilizam algoritmos que intentam despertar as emoções de seu parceiro. Nesse sentido, o robô sexual também passa a armazenar e processar informações íntimas do usuário, o que requer garantias para que tais dados permaneçam privados e seguros contra ataques de *hackers.*

A preocupação acerca da relação entre o fluxo de dados e a inteligência artificial é ressaltada por Eduardo Tomasevicius Filho[28],

> quando computadores estavam isolados uns dos outros, a capacidade da inteligência artificial limitava-se aos dados disponíveis nas memórias dessas maquinas. Porém, com a melhoria dos *softwares* de reconhecimento de textos, imagens e informações originalmente registradas em suportes materiais – e, sobretudo, com a possibilidade de acesso a esses dados de maneira instantânea em qualquer parte do mundo por meio da Internet, além do armazenamento de informações em grandes servidores de dados, também conhecida como "computação na nuvem" –, a inteligência artificial assumiu nova dimensão, porque possibilitou o acesso a informações *ad infinitum.* Além

[27] A Lei nº 13.709/18 teve sua *vacatio legis* alterada por força da Lei nº 13.853/19, de modo que entrará em vigor no dia 28 de dezembro de 2018, quanto aos arts. 55-A, 55-B, 55-C, 55-D, 55-E, 55-F, 55-G, 55-H, 55-I, 55-J, 55-K, 55-L, 58-A e 58-B e 24 (vinte e quatro) meses após a data de sua publicação (15.8.2018), quanto aos demais artigos.

[28] TOMASEVICIUS FILHO, Eduardo. Inteligência artificial e direitos da personalidade: uma contradição em termos? *Revista da Faculdade de Direito da Universidade de São Paulo*, São Paulo, v. 113, p. 133-149, jan./dez. 2018. p. 137.

> disso, a Internet facilita a formação e coleta de *big data,* isto é, de informações relativas à navegação pela rede, como também sobre o que é inserido ou consultado pelo interessado.

Também se ressalta a necessidade de respeito pelo regime democrático, de modo que os sistemas de inteligência artificial devem respeitar a pluralidade de valores e escolhas individuais, sem prejudicar compromissos fundamentais sobre os quais o Estado Democrático de Direito se fundamenta, tais como a isonomia e o devido processo legal.

Tal preocupação é especialmente relevante quando se constata que a inteligência artificial vai muito além das recomendações que são apresentadas no *Spotify* ou na *Netflix*. Evidencia-se a crescente utilização da inteligência artificial por órgãos judiciais, o que vem se verificando, inclusive, no Brasil. No Superior Tribunal de Justiça[29], por exemplo, desenvolveu-se um projeto-piloto que, na Secretaria Judiciária, automatizará a definição do assunto do processo na classificação processual e na extração automática de dispositivos legais apontados como violados. No Supremo Tribunal Federal desenvolve-se um sistema chamado VICTOR, cujo objetivo é ler os recursos extraordinários que chegam a essa Corte e identificar a vinculação com determinados temas de repercussão geral.

Também são conhecidas as chamadas *startups law techs*[30], que desenvolvem "robôs advogados" capazes de auxiliar o profissional na coleta de dados, organização de documentos, cálculos, formatação, intepretações judiciais, prognósticos de decisões etc. Para além de tais funções, nos EUA, algoritmos de avaliação de risco vêm sendo utilizados para medir a probabilidade de reincidência de um acusado[31], em fenômeno chamado de *predictive justice.* Nesse sentido, o americano Eric Loomis foi condenado a seis anos de prisão, com base numa previsão algorítmica secreta que concluiu que o sujeito voltaria a cometer crimes[32].

Indaga-se se tal sistema seria compatível com a axiologia constitucional brasileira, considerando que tal fundamentação afeta inequivocamente os direitos do contraditório e da ampla defesa, pilares do devido processo legal num regime democrático. É incontestável que os sistemas de IA possuem ampla capacidade de melhorar a eficiência da prestação dos serviços públicos. Ocorre que, paradoxalmente, também possuem capacidade de afetar negativamente os direitos fundamentais dos cidadãos. O modo do exercício punitivo estatal sobre seus cidadãos é uma manifestação direta da caracterização do sistema político. O ente estatal pode se manifestar como autoritário ou democrático a depender da maneira como estrutura sua relação com o meio social.

---

[29] Disponível em: http://www.stj.jus.br/sites/STJ/default/pt_BR/Comunicação/noticias/Notícias/STJ-d%C3%A1-primeiro-passo-para-implantar-inteligência-artificial-na-rotina-do-processo Acesso em: 28 set. 2018.

[30] Disponível em: https://www.infomoney.com.br/negocios/inovacao/noticia/6757258/primeiro-robo-advogado-brasil-lancado-por-empresa-brasileira-conheca Acesso em: 28 set. 2018.

[31] Disponível em: http://parisinnovationreview.com/articles-en/predictive-justice-when-algorithms-pervade-the-law. Acesso em: 14 maio 2019.

[32] Disponível em: https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html?_r=0. Acesso em: 14 maio 2019.

Ainda na questão criminal, um relatório da ProPublica[33] indicou que os algoritmos expunham vieses racistas na aplicação da lei. A fórmula culminava por denunciar equivocadamente réus negros como futuros criminosos, rotulando-os quase duas vezes mais como criminosos de alto risco, mesmo quando não reincidiam de fato. A empresa responsável pelo desenvolvimento do sistema refutou as acusações e aduziu que as conclusões foram extraídas por meio de um questionário de 137 perguntas respondidas pelos réus ou extraídas de registros criminais.

É imprescindível que a utilização da IA no âmbito judicial ocorra de forma transparente, mormente tendo em vista o princípio da publicidade na Administração Pública, estampado no art. 37 da Constituição Federal. No mesmo sentido, só é possível questionar os fundamentos de uma decisão automatizada quando se conhecem os critérios previamente estipulados. Não se ignoram, contudo, as dificuldades que podem surgir em face da propriedade intelectual do programador. Seria igualmente desejável que a autoridade responsável pela custódia de tais dados os tratasse com sigilo, bem como todas as partes envolvidas na verificação das questões que se fizessem necessárias.

É fundamental que também haja uma postura de ceticismo acerca da concepção de neutralidade dos dados. Isso porque a inteligência artificial se baseia numa grande quantidade de dados e informações cuja mineração depende, sobretudo, de escolhas dos programadores. A operação depende essencialmente de *inputs* e de *outputs* do programador.

Se os dados subjacentes são tendenciosos, as desigualdades estruturais e os preconceitos inculcados nos dados serão amplificados por meio da atividade da inteligência artificial. As próprias escolhas sobre inserção, organização e classificação de dados deve ser feita de modo cauteloso por todos os envolvidos, sob pena de violação aos direitos de personalidade.

A dificuldade exsurge quando se constata que tais dados são reflexos de problemas estruturais verificados na sociedade. Por exemplo, quando se argumenta que existem evidências de que americanos negros são presos cerca de quatro vezes mais que os americanos brancos[34]. Se houver a coleta fiel de tais dados pelo algoritmo, a IA incorpora e reflete esse viés.

É possível que surjam simplificações inadequadas em face de situações sociais complexas que exigem um raciocínio mais aprofundado, o que exige um papel proativo e cauteloso do programador, que busque assegurar ampla representação nos dados, para que seja possível reduzir distorções e assegurar condições imparciais.

A própria noção de julgamento por uma máquina também recebe críticas. Argumenta-se que

> permitir que uma máquina tome determinada decisão em âmbito jurisdicional só seria possível se se concebesse o processo jurisdicional como uma mera

[33] Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 14 maio 2019.
[34] Disponível em: https://www.politifact.com/punditfact/statements/2016/jul/13/van-jones/van-jones-claim-drug-use-imprisonment-rates-blacks/ Acesso em: 14 maio 2019.

> escolha entre as várias disponíveis, e sem que se considerasse a importância da hermenêutica e dos valores (éticos, sociais e morais) para tal processo.[35]

Evidencia-se, nesse ponto, a necessidade de abertura do sistema jurídico para argumentos pragmáticos e éticos, desvinculando-se de uma perspectiva hermética que se funda em dados limitados, para que se possa assegurar um efetivo diálogo democrático.

As Diretrizes também apontam a necessidade de observar a igualdade, a não discriminação e a solidariedade, especialmente os direitos das pessoas em situação de vulnerabilidade ou exclusão. Nesse contexto, a isonomia implica precisamente a impossibilidade de que os sistemas desenvolvam situações tendenciosas, devendo ser inclusivos e representar grupos populacionais diferentes, com especial atenção a grupos tidos como vulneráveis.

No item 2.2, o documento também elenca princípios a serem observados no desenvolvimento da inteligência artificial. Nesse sentido, elenca-se o princípio do respeito pela autonomia humana, de modo que "os seres humanos que interagem com os sistemas de IA devem ser capazes de manter uma autodeterminação plena e efetiva sobre si mesmos e poder participar do processo democrático". [36] - [37] Não poderia haver, portanto, subordinação ou manipulação dos seres humanos por meio da inteligência artificial, devendo esta servir para complementar e fomentar as habilidades cognitivas, sociais e culturais dos agentes, deixando margem de escolha ao ser humano.

Tal perspectiva implica assegurar a supervisão humana acerca do funcionamento da inteligência artificial, o que também enseja ponderações sobre noções de IA forte [38], considerando que em tais situações a IA possui maior capacidade de autossuficiência.

Outro princípio elencado é o da prevenção do dano. Esse princípio determina que os sistemas de IA não devem causar nem agravar danos ou, de outra forma, afetar adversamente os seres humanos. Trata-se da consagração da incolumidade das esferas jurídicas, fundamentada na dignidade e na integridade mental e física do ser humano. Como corolário, torna-se imprescindível que os ambientes de operação sejam suficientemente seguros e tecnicamente robustos, dando especial atenção às situações em que possam existir vulnerabilidades e assimetrias de poder ou informação, em consideração ao ambiente natural de todos os seres humanos.

A prevenção dos danos é um imperativo cada vez mais constante na contemporânea sociedade de risco. Diariamente surgem notícias acerca de ataques de

---

[35] OLIVEIRA, Samuel Rodrigues De; COSTA, Ramon Silva. Pode a máquina julgar? Considerações sobre o uso de inteligência artificial no processo de decisão judicial. *Revista de Argumentação e Hermenêutica Jurídica*, Porto Alegre, v. 4, n. 2, p. 21-39, jul./dez 2018. p. 11.

[36] "Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process". Ethics Guidelines for trustworthy AI. Disponível em: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Acesso em: 4 maio 2019, p. 12.

[37] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINE. *Ethics Guidelines for trustworthy AI.* Disponível em: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Acesso em: 4 maio 2019. p. 12.

[38] STRELKOVA, O. PASICHNYK, O. *Three types of artificial intelligence.* Disponível em: http://eztuir.ztu.edu.ua/jspui/bitstream/123456789/6479/1/142.pdf. Acesso em: 3 maio 2019.

hackers[39] ou vazamentos indevidos de dados[40], o que seguramente tem o condão de violar direitos de personalidade dos usuários. Com efeito, sob a perspectiva de Ulrich Beck em sua obra "A sociedade de risco", a sociedade contemporânea é marcada por perigos que se situam na imbricação entre construções científicas e sociais, sendo o desenvolvimento tecnológico uma fonte de causa, definição e solução de riscos. A IA deve assumir, nesse contexto, protagonismo na tentativa de mitigação e gerenciamento de crises.

As Diretrizes também referem o princípio da justiça, suscitando que tal princípio deve possuir dimensão substantiva e processual. A dimensão substantiva implica o compromisso de distribuição igualitária e equânime de benefícios e custos, bem como a ausência de discriminações e estigmatizações, na observância da proporcionalidade entre fins e meios e no equilíbrio entre objetivos concorrentes.

Noutro norte, na dimensão processual, ressalta-se a capacidade de contestar e buscar reparações efetivas em face de lesões causadas por mecanismos de IA, bem como pelos agentes que os operam. Para tanto, argumenta-se que "a entidade responsável pela decisão deve ser identificável, e os processos de tomada de decisão devem ser explicáveis[41]".

Exsurge, nesse ponto, a necessidade de que os algoritmos sejam auditáveis e que a tomada de decisão possa ser compreendida pelos interessados. Simultaneamente, deve-se assegurar que os interesses empresariais do programador não sejam comprometidos, o que, por si só, já cria desafios jurídicos a serem desenvolvidos pela doutrina e pelo legislador.

Compreende-se, ainda, que a terminologia "princípio da justiça"[42] não foi a mais adequada, mormente considerando a multiplicidade de conceitos e interpretações que tal vocábulo pode denotar. Analisando a ideia de justiça desenvolvida pelas Diretrizes, pode-se concluir que se quis fazer uma amálgama em referência a princípios de equidade, liberdade, proporcionalidade, solidariedade social e devido processo legal.

As Diretrizes também indicam o princípio da explicabilidade[43]. Esta possui o objetivo de manter a transparência e a confiança dos usuários na tecnologia, devendo expor as

---

[39] Disponível em: https://g1.globo.com/economia/tecnologia/noticia/2019/05/14/whatsapp-detecta-vulnerabilidade-que-permite-o-acesso-de-hackers-a-celulares.ghtml. Acesso em: 12 maio 2019.

[40] Disponível em: http://s.migalhas.com.br/S/53FC0B. Acesso em: 13 maio 2019.

[41] "The entity accountable for the decision must be identifiable, and the decision-making processes should be explicable".

[42] "The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatization. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable."

[43] "This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not

capacidades e o propósito do sistema de IA a todos aqueles que sejam direta ou indiretamente afetados. Numa sociedade marcada pelo consumo, o Código de Defesa do Consumidor traz como direito básico, em seu art. 6º, III, a informação adequada e clara sobre os diferentes produtos e serviços, incluindo os riscos que apresentem.

Ressalte-se que o dever de informação pode encontrar dificuldades, pois, muitas vezes, não há conhecimento exato do funcionamento da IA e de suas potencialidades. Especialmente em situações de utilização de algoritmos por órgãos judiciais, a explicabilidade assume grande relevância, porquanto uma decisão não pode ser efetivamente contestada sem que se possa auditar a inteligência artificial e compreender, em termos mínimos, o seu funcionamento.

Não se ignora que a existência da *black box* da IA, compreendida como a ausência de conhecimento acerca da tomada de uma saída ou decisão específica (ou de quais fatores contribuíram para isso), pode desencadear dificuldades. Ainda assim, o respeito aos direitos fundamentais exige que se adote uma postura de explicabilidade, com rastreabilidade e comunicação transparente acerca das capacidades e limitações conhecidas do sistema até então. Trata-se de uma análise contextual, que verifica o funcionamento, os dados, o estado da arte e os objetivos usualmente visados pelo mecanismo tecnológico.

Em síntese, as Diretrizes estabelecem os princípios do respeito pela autonomia humana, da prevenção do dano, da justiça e da explicabilidade. Inevitavelmente, tensões podem surgir da interação entre tais princípios, bem como com outros princípios e direitos fundamentais elencados pelo ordenamento jurídico brasileiro, não havendo uma resposta apriorística e fixa para a resolução de tais impasses.

Nesse ponto, é salutar a lição de Alexy[44] (2015, p. 588) no sentido de que os princípios são normas que ordenam que algo seja realizado na maior medida possível dentro das possibilidades jurídicas e fáticas existentes. Trata-se de analisar os fatores do contexto e ponderar quais interesses devem prevalecer. Em observância ao desenvolvimento do Estado Democrático de Direito é imprescindível que a percepção e a resolução de tais tensões sejam feitas de modo dialógico com todos os interessados.

Políticas públicas que utilizam reconhecimento facial[45] desencadeiam profundos debates acerca de conflitos entre a noção de justiça, autonomia humana, privacidade e os deveres estatais de proteção e segurança.[46-47] Ainda sob o prisma de Alexy[48], uma restrição a

---

always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate".

[44] ALEXY, Robert. *Teoria dos direitos fundamentais*. Tradução de Virgílio Afonso da Silva. 2. ed. São Paulo: Malheiros, 2015. p. 588.

[45] Disponível em: https://globoplay.globo.com/v/7619379/. Acesso em: 14 maio. 2019.

[46] Confira-se: ALEXY, Robert. *Teoria dos direitos fundamentais*. Tradução de Virgílio Afonso da Silva. 2. ed. São Paulo: Malheiros, 2015. p. 588.

[47] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINE. Ethics Guidelines for trustworthy AI. Disponível em: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Acesso em: 4 maio 2019.

um direito fundamental somente é admissível se, no caso concreto, aos princípios colidentes for atribuído um peso maior que aquele atribuído ao princípio de direito fundamental em questão, de modo que a solução de tal conflito perpassa pela ponderação e atribuição de pesos aos interesses.

Impende ressaltar que as diretrizes não propõem soluções definitivas e prontas para as questões do desenvolvimento tecnológico, caracterizando-se como prescrições éticas abstratas para todos os interessados, que serão utilizadas por meio de reflexões fundamentadas como norte para a atividade empresarial e para a resolução de impasses e que ainda carecem de adaptação para a realidade socioeconômica e jurídica brasileira.

4 DOS REQUISITOS DE CONFIABILIDADE

A partir do segundo capítulo das Diretrizes, elencam-se de modo exemplificativo os requisitos que devem ser observados para que o desenvolvimento da inteligência artificial seja confiável: a) agência e fiscalização humana; b) robustez e segurança; c) privacidade e governança de dados; d) transparência; e) diversidade, não discriminação e equidade; f) bem-estar social e ambiental; g) e responsabilização.

A verificação de tais requisitos demanda que haja pesquisa acerca dos sistemas de IA, com divulgação de resultados e abertura de questões ao público. O primeiro requisito reporta-se ao princípio da autonomia humana e requer que os sistemas de IA apoiem a tomada de decisões e permitam a supervisão humana sobre seu funcionamento, reafirmando o compromisso antropocêntrico.

Parte-se da perspectiva de que a tecnologia é um paradoxo, ao passo que simultaneamente é fator de causa e solução de riscos para os direitos fundamentais. Nesse sentido, a tecnologia pode ajudar indivíduos a terem maiores chances de cura de uma patologia ou aumentar a acessibilidade em educação para pessoas com deficiência, privilegiando o direito à saúde e à educação. Ao mesmo tempo, também pode violar a privacidade dos indivíduos e causar prejuízos imprevisíveis.

É imprescindível, portanto, que em tais situações haja uma avaliação dos impactos e uma tentativa de mitigação dos riscos necessários em uma sociedade democrática, com vistas a não infringir as esferas jurídicas dos indivíduos. No mesmo sentido, é importante que haja mecanismos de *feedback* externo sobre os sistemas de IA, com vistas a proporcionar um funcionamento dialógico com a sociedade.

Ainda tratando do mesmo requisito, é importante que os usuários sejam capazes de tomar decisões informadas sobre os sistemas de IA. Sobre o direito à informação, Paulo Lôbo[49]

---

[48] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINE. *Ethics Guidelines for trustworthy AI.* Disponível em: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Acesso em: 4 maio 2019. p. 12. ALEXY, Robert. *Teoria dos direitos fundamentais*. Tradução de Virgílio Afonso da Silva. 2. ed. São Paulo: Malheiros, 2015. p. 588.

[49] LÔBO, Paulo. A informação como direito fundamental do consumidor. *Revista de Direito do Consumidor*,

(2001, p. 66) argumenta ser um direito fundamental que se concretiza quando a informação recebida pelo consumidor típico preenche os requisitos de adequação, suficiência e veracidade. A adequação diz respeito aos meios de informação utilizados e seus respectivos conteúdos e linguagens; a suficiência diz respeito à completude e integralidade da informação; e, por fim, a veracidade diz respeito às reais características do produto ou serviço, além dos dados corretos acerca da composição, conteúdo, preço, prazos, garantias e riscos. O direito fundamental à informação concretiza-se quando o consumidor tem capacidade objetiva de conhecimento e compreensão acerca daquele produto ou serviço.

Ainda nesse ponto, é crucial a perspectiva de Paulo Lôbo no sentido de que "insuficiente é, também, a informação que reduz, de modo proposital, as consequências danosas pelo uso do produto, em virtude do estágio ainda incerto do conhecimento científico ou tecnológico". No campo da inteligência artificial, incumbe aos desenvolvedores informar o consumidor acerca da capacidade da máquina conhecida até então, advertindo-o da incerteza ou obscuridade no que tange a alguns funcionamentos da tecnologia.

Impende evidenciar a questão do direito de não estar sujeito a uma decisão baseada unicamente no processamento automatizado quando existirem efeitos legais sobre os usuários ou especificidades do caso concreto que demandem uma análise mais acurada. A supervisão humana deve ser garantida, ajudando a eliminar eventuais distorções.

O requisito da robustez e segurança, por sua vez, inclui a proteção aos ataques de *hackers*, planos de retorno e confiabilidade. A robustez técnica está intimamente ligada ao princípio da prevenção de danos, porquanto determina que os sistemas possuam abordagem preventiva de riscos e que se comportem de modo confiável, minimizando a ocorrência de danos inesperados. É imprescindível, portanto, que se desenvolvam proteções contra vulnerabilidades, máxime considerando a amplitude e influência que o *hacking* pode ter sobre o funcionamento da máquina. Medidas de segurança insuficientes também podem ensejar decisões equivocadas ou até danos físicos ao usuário.

O nível de medidas de segurança requeridas em determinado sistema de IA deverá ser proporcional ao nível da magnitude do risco apresentado pela máquina. O risco será maior na medida em que a máquina tiver menor precisão em seu funcionamento, isto é, menor capacidade de fazer julgamentos e classificações corretas com base nos dados ou modelos.

As Diretrizes estipulam, ainda, que "quando previsões imprecisas ocasionais não puderem ser evitadas, é importante que o sistema possa indicar a probabilidade de ocorrência desses erros[50]", o que concretiza a noção de boa-fé objetiva. É de extrema relevância que haja um mínimo controle sobre aquilo que se produz, para que seja possível, de modo transparente, informar os interessados sobre as limitações verificadas no produto ou serviço.

É fundamental que os resultados dos sistemas de IA sejam reprodutíveis e confiáveis, isto é, que funcionem adequadamente numa variedade de situações e de modo

São Paulo: Revista dos Tribunais, ano 10, n. 37, 2001. p. 66.

[50] "When occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are".

similar quando repetidos sob as mesmas condições, o que contribuirá para a obtenção do nível de precisão do sistema.

No que tange ao requisito da privacidade e proteção de dados, as Diretrizes indicam que o sistema deve garantir a integridade de todas as informações inicialmente fornecidas pelo usuário, bem como as informações geradas durante a interação com o sistema, assegurando que os dados coletados não sejam utilizados indevidamente. No ordenamento brasileiro, trata-se de assegurar a observância do art. 5º, X, da Constituição Federal.

No contexto contemporâneo, a privacidade abandona a clássica concepção de ser o mero "direito de estar só", para abranger outras facetas de controle sobre as informações pessoais, especialmente na sociedade digital. Stefano Rodotà[51] desenvolve a concepção de autodeterminação informativa como direito fundamental e argumenta que

> coerentemente com a mudança da própria definição de privacidade, a atenção deve passar do sigilo ao controle. Isto significa, em primeiro lugar, que se torna cada vez mais difícil individuar tipos de informações acerca dos quais o cidadão estaria disposto a "despir-se" completamente, no sentido de renunciar definitivamente a controlar as modalidades de seu tratamento e a atividade dos sujeitos que a utilizam. Esta concepção depende sobretudo da percepção de que até as informações aparentemente mais inócuas podem, se integradas a outras, provocar dano ao interessado. E não se pode dizer que tal comportamento esteja em contradição com a tendência, anteriormente referida, segundo a qual existem categorias inteiras de informações pessoais (como aquelas de conteúdo econômico) cuja divulgação é oportuna ou necessária: publicidade e controle não são termos contraditórios, como são publicidade e sigilo. Exatamente onde se admitir a máxima circulação das informações de conteúdo econômico, deve-se permitir aos interessados exercitar um real poder de controle sobre a exatidão de tais informações, sobre os sujeitos que as operam e sobre as modalidades de sua utilização. Em segundo lugar, e sobretudo, a nova situação determinada pelo uso de computadores no tratamento das informações pessoais torna cada vez mais difícil considerar o cidadão como um simples "fornecedor de dados", sem que a ele caiba algum poder de controle. De fato, a obrigação de fornecer dados não pode ser simplesmente considerada como a contrapartida dos benefícios sociais que, direta ou indiretamente, o cidadão pode chegar a aproveitar. As informações coletadas não somente tornam as organizações públicas e privadas capazes de planejar e executar os seus programas, mas permitem o surgimento de novas concentrações de poder ou o fortalecimento de poderes já existentes: consequentemente, os cidadãos têm o direito de pretender exercer um controle direto sobre aqueles sujeitos aos quais as informações fornecidas atribuirão um crescente plus-poder.

Nesse sentido, impõe-se o respeito pela privacidade, qualidade, integridade e acesso aos dados. Nos termos de Erick Lucena e Marcos Ehrhardt Júnior[52], algumas estratégias podem ser traçadas para a tutela da privacidade:

> a primeira delas seria o "direito de oposição", que, de forma individual ou coletiva, funcionaria como uma negativa à coleta e circulação de informações pessoais em determinadas formas. b) O "direito de não saber" é a segunda

[51] RODOTÀ, Stefano. *A vida na sociedade de vigilância*: a privacidade hoje. Tradução de Danilo Doneda e Luciana Cabral Doneda. Rio de Janeiro: Renovar, 2008. p. 66.
[52] PEIXOTO, Erick Lucena Campos; EHRHARDT JÚNIOR, Marcos. Breves notas sobre a ressignificação da privacidade. *Revista Brasileira de Direito Civil – RBDCivil*, Belo Horizonte, v. 16, p. 35-56, abr./jun. 2018. p. 44.

> estratégia de tutela da privacidade, podendo ser tratado como decorrente do primeiro. Surgido em relação a dados de saúde, passa a ser estendido contra as formas de marketing direto que invadem a esfera privada do indivíduo com informações não solicitadas e não desejadas. c) Outra estratégia é tornar mais clara a finalidade da coleta de dados. A legitimidade aqui é condicionada à comunicação preventiva ao interessado sobre o motivo da coleta e o destino dos dados coletados. d) Por último, o "direito ao esquecimento", "prevendo-se que algumas categorias de informações devam ser destruídas, ou conservadas somente em forma agregada e anônima, uma vez que tenha sido atingida a finalidade para a qual foram coletadas" ou ainda, "depois de transcorrido um determinado lapso de tempo.

A qualidade e a integridade dos dados são cruciais ao desempenho dos sistemas de IA, considerando que quando os dados coletados contêm vieses, imprecisões ou falhas estruturais, tais erros serão reproduzidos pela tecnologia. No mesmo sentido, a alimentação de dados maliciosos pode mudar o comportamento do sistema, especialmente no que tange à autoaprendizagem.

Por tais motivos, é importante que o conjunto e o processamento dos dados sejam testados e documentados em cada etapa, desde o planejamento, o treinamento, o teste, até a implantação. Ademais, é interessante que haja um protocolo interno que discipline o acesso aos dados e suas respectivas circunstâncias, de modo que somente pessoas interessadas e qualificadas possam acessá-los. Ressalte-se que a proteção de dados pessoais está prevista no art. 8º[53] da Carta de Direitos Fundamentais da União Europeia, e que a Diretiva 95/46/CE trata especificamente sobre as definições de dados pessoais e arquivos de dados pessoais, além de fixar princípios específicos.

O requisito da transparência, por sua vez, relaciona-se com o princípio da explicabilidade e abrange os elementos relevantes para um sistema de IA, quais sejam: os dados, o sistema e os modelos de negócio, exigindo rastreabilidade, explicabilidade e comunicação.

A relevância da rastreabilidade se perfaz quando o conjunto de dados e processos que geram a decisão do sistema IA, incluindo os de coleta e classificação de dados, bem como os algoritmos usados, são documentados. Isso torna mais fácil a compreensão das razões pelas quais uma decisão da IA foi tomada, o que ajuda a mitigar a ocorrência de erros futuros.

A explicabilidade diz respeito à capacidade de explicar os processos técnicos de tomada de decisão do sistema de inteligência artificial. Quanto maior for o impacto da tomada de decisão na vida do indivíduo, maior deve ser a possibilidade de exigir uma explicação acerca do processo, devendo tal explicação ser oportuna e adequada à perícia do *stakeholder* envolvido.

A transparência no modelo de negócios diz respeito às explicações de quanto um sistema de inteligência artificial influencia no processo de tomada de decisões organizacionais e nas escolhas dos projetos, bem como qual a justificativa para implantá-los.

A comunicação refere-se à necessidade de que o usuário seja previamente advertido que estará interagindo com um sistema de IA, de modo que a natureza da máquina

[53] "Art. 8º 1. Todas as pessoas têm direito à proteção dos dados de caráter pessoal que lhes digam respeito."

seja identificável, bem como suas eventuais limitações e capacidades. Ressalte-se que a contribuição mais famosa de Alan Turing para a inteligência artificial foi o "Experimento mental", também conhecido como Teste de Turing[54], que se caracteriza quando um ser humano se comunica com uma parte desconhecida, que poderia ser uma pessoa ou um computador. Se o computador fosse capaz de responder ao ser humano de modo que este acreditasse que se tratava de uma pessoa e não de uma máquina, haveria fortes evidências de que o computador realmente era inteligente.

Nesse sentido, a inteligência artificial se caracterizaria por mimetizar a experiência humana de modo a fazer o interlocutor acreditar que se tratava de uma pessoa. O elemento da comunicação, nesse ponto, serve como obstáculo a essa confusão, de modo que o usuário deve ser previamente advertido de que está fazendo uso de uma tecnologia artificial.

O requisito da diversidade, não discriminação e justiça diz respeito à inclusão e à diversidade no processo de IA, envolvendo todos os *stakeholders* afetados ao longo do processo, bem como garantindo igualdade de acesso aos interessados. Nesse sentido, é imprescindível que sejam avaliados os critérios usados pelos sistemas de IA, uma vez que podem sofrer inclusões de modelos com vieses inadequados que podem ensejar preconceitos e discriminações não intencionais contra certos grupos, exacerbando problemas estruturais de marginalização.

Os vieses discriminatórios devem ser tolhidos já na fase de coleta, de modo que os critérios a serem utilizados no processamento da IA já estejam livres de tais falhas. É importante, assim, que a base de dados seja inclusiva no que tange a diversas culturas e origens. Tais problemas também podem ser mitigados com supervisões que analisem finalidade, restrições, requisitos e decisões do sistema de maneira coerente e transparente.

> É fácil perceber que, se forem utilizados no modelo estatístico dados com alto potencial discriminatório, tais como dados raciais, étnicos ou de orientação sexual, haverá um grande risco de que a decisão que resultará do processo automatizado (*output*) também seja discriminatória. Esses dados são os chamados dados sensíveis, cujo processamento é limitado pelas legislações de proteção de dados de vários países, assim como pelo Regulamento Europeu de Dados Pessoais. Em segundo lugar, é preciso observar que o próprio método utilizado nas decisões automatizadas – por meio da classificação e seleção dos indivíduos – gera um risco de se produzirem resultados discriminatórios, ainda que de forma não intencional. Isto pode ocorrer porque, na discriminação estatística, teoria econômica que se tornou conhecida a partir dos textos de Edmund Phelps (1972) e Kenneth Arrow (1973), os indivíduos são diferenciados com base em características prováveis de um grupo, no qual esse indivíduo é classificado. Essa prática se baseia em métodos estatísticos, que associam esses atributos a outras características, cuja identificação pelo tomador de decisão é mais difícil, como nível de renda, risco de inadimplência, produtividade no trabalho, etc. (BRITZ, 2008, p. 15). Nesse contexto, é possível a ocorrência da discriminação por erro estatístico, o que decorreria tanto de dados incorretamente capturados como também de modelo estatístico de bases científicas frágeis (BRITZ, 2008).[55]

[54] HENDERSON, Harry. *Artificial intelligence*: mirrors for the mind. New York: Chelsea House Publishers, 2007. p. 29.
[55] DONEDA, Danilo Cesar Maganhoto; MENDES, Laura Schertel; SOUZA, Carlos Affonso Pereira de; ANDRADE, Norberto Nuno Gomes de. Considerações iniciais sobre inteligência artificial, ética e autonomia

Além disso, podem surgir problemas de discriminação por outros meios: “resultados discriminatórios também são possíveis por meio da generalização, prática muito utilizada nas decisões automatizadas, o que levou Gabriele Britz (...) a cunhar a expressão “injustiça pela generalização”. A discriminação estatística se dá por meio da classificação de pessoas com determinadas características em certos grupos – isto é, por meio da generalização de que pessoas com tais características têm maior probabilidade de agir de certa maneira ou de apresentar determinadas qualidades. A generalização, nesse caso, embora o modelo possa funcionar bem e seja estatisticamente correto, pode levar à discriminação das pessoas que configuram os casos atípicos, não se enquadrando nas características do grupo geral. É o caso, por exemplo, da pessoa que, apesar de morar em determinada região, considerada de baixa renda e, portanto, classificada como de maior risco de inadimplência em modelos de risco de crédito, aufere na realidade renda superior à de seus vizinhos. A discriminação, nesse caso, dar-se-ia, porque, em um modelo em que a informação sobre endereço tem peso fundamental, o caso atípico seria tratado conforme o grupo em que está inserido, e não conforme as outras pessoas de sua faixa de renda.”[56]

As Diretrizes ressaltam, ainda, que deve haver padrão de acessibilidade e *design* universal. Isto é, os sistemas devem ser centrados no usuário e projetados de forma a permitir que todas as pessoas utilizem produtos ou serviços de IA, independentemente de idade, sexo, habilidades ou características, especialmente pessoas vulneráveis.

Recomenda-se que os *stakeholders* sejam consultados durante o ciclo de vida do sistema, solicitando *feedback* e estabelecendo mecanismos de longo prazo para a participação das partes, a ensejar um ambiente dialógico e democrático. Assim, José Barros Correia Júnior trata dos *stakeholders* sob a perspectiva de que a empresa é uma atividade concentradora de interesses múltiplos, indo além do tradicional negócio de interesses exclusivos dos investidores[57]

Sob a perspectiva da função social e da responsabilidade social, a teoria do *stakeholders* parte do pressuposto de que a empresa se caracteriza pela confluência de interesses múltiplos, que vão além do lucro do investidor e com ele se relacionam, sendo imprescindível considerar as atividades das partes interessadas.

O requisito do bem-estar social e ambiental refere-se ao desenvolvimento sustentável, respeito pelo meio ambiente e impacto social. Nesse ponto, o art. 225 da Constituição Federal brasileira estipula que todos têm direito ao meio ambiente ecologicamente equilibrado, bem de uso comum do povo e essencial à sadia qualidade de vida, impondo-se ao Poder Público e à coletividade o dever de defendê-lo e preservá-lo para as presentes e futuras gerações. Trata-se, portanto, de direito fundamental baseado na noção de solidariedade social.

---

pessoal. *Pensar: Revista* de Ciências Jurídicas, Fortaleza, v. 23, n. 4, p. 1-17, out./dez. 2018. p. 5.

[56] DONEDA, Danilo Cesar Maganhoto; MENDES, Laura Schertel; SOUZA, Carlos Affonso Pereira de; ANDRADE, Norberto Nuno Gomes de. Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal, cit., p. 5.

[57] CORREIA JÚNIOR, José Barros. *A função social e a responsabilidade social da empresa perante os stakeholders.* 2013. Tese (Doutorado em Direito). Pós-Graduação em Direito da Faculdade de Direito de Recife da Universidade Federal de Pernambuco, Recife. p. 173.

Os desenvolvedores da tecnologia devem lidar com as preocupações ambientais, sem negligenciar os impactos que podem advir de sua atuação. O processo de desenvolvimento, bem como toda a sua cadeia de produção e suprimentos, deve ser avaliado por meio da análise de seus recursos e consumo de energia, optando sempre por opções menos prejudiciais. Trata-se da consagração da função social da empresa, que tem o dever de observar as normas cogentes que versam sobre a preservação do meio ambiente.

Também se recomenda que os efeitos sociais oriundos da IA sejam devidamente analisados e monitorados, considerando que a tecnologia se torna cada vez mais presente e invasiva no cotidiano dos indivíduos, às vezes de modo bem sutil. Tal fenômeno pode ensejar modificações nas relações sociais, econômicas e culturais, contribuindo e simultaneamente deteriorando habilidades e costumes sociais. O uso de sistemas de IA deve ser cautelosamente analisado, máxime tendo em vista que até mesmo em contextos eleitorais[58] a tecnologia vem sendo utilizada para influenciar posturas e perspectivas sociais.

Por fim, o requisito da responsabilização se refere à possibilidade de auditoria, minimização de impactos negativos e reparação. A auditoria se refere à avaliação de algoritmos, dados e processos de *design.* Muito se argumenta acerca da possível violação de modelos de negócios e propriedade intelectual dos programadores por auditores externos e internos, o que demanda esforços no sentido de compatibilizar o sigilo de tais empresários com a necessidade de auditoria independente. Em sistemas que afetam direitos fundamentais, incluindo aplicações críticas de segurança, é imprescindível que os sistemas possam ser auditados.

É também importante que sejam relatadas as ações e decisões que contribuem para um determinado resultado do sistema, identificando e avaliando os potenciais impactos negativos dos sistemas de IA. Quanto maior for o risco que o sistema apresente, maior deve ser a cautela com as avaliações. A reparação também é um aspecto de relevância em face do princípio da reparação integral da vítima no ordenamento pátrio, de modo que devem existir mecanismos acessíveis que garantam reparação adequada àqueles que suportarem algum prejuízo oriundo do desenvolvimento da tecnologia.

A necessidade de reparação se avulta sob a noção da teoria do risco, com fulcro no art. 927, parágrafo único, do Código Civil, que determina que há obrigação de reparar o dano independentemente de culpa, nos casos especificados em lei, ou quando a atividade normalmente desenvolvida pelo autor do dano implicar, por sua natureza, risco para os direitos de outrem. Trata-se da responsabilidade objetiva, que dispensa verificação de culpa para incidência e se lastreia na necessidade de assegurar à vítima a reparação de seu prejuízo.

Nesse sentido, argumenta Maria Celina Bodin de Moraes:

> Com o passar do tempo, porém, o dever de solidariedade social, o fundamento constitucional da responsabilidade objetiva, sobressairá e aceitar-se-á que seu alcance é amplo o suficiente para abranger a reparação de todos os danos injustamente sofridos, em havendo nexo de causalidade com a atividade

[58] Disponível em: https://g1.globo.com/economia/tecnologia/noticia/o-que-e-um-robo-na-web-e-como-ele-pode-influenciar-o-debate-nas-redes-especialistas-explicam.ghtml. Acesso em: 21 maio 2019.

> desenvolvida, seja ela perigosa ou não. Não se sustentará mais qualquer resquício de culpa, de sanção ou de descumprimento de deveres no fundamento da responsabilidade objetiva. Com efeito, todas são atividades que geram 'risco para os direitos de outrem', como prevê o dispositivo legal.[59]

A preocupação com a responsabilidade objetiva é especialmente relevante no contexto contemporâneo de tutela de vulneráveis. A relevância dessa tutela tem como fundamento o conceito de isonomia material, que parte da constatação da desigualdade fática e demanda mecanismos de compensação para que a todos seja atribuída a mesma possibilidade de direitos.

5 PARÂMETROS DE AVALIAÇÃO DA INTELIGÊNCIA ARTIFICIAL

Por fim, as Diretrizes trazem uma série de requisitos que podem ser utilizados como parâmetro de avaliação para a confiabilidade da inteligência artificial. Nesse ponto, impende verificar, por exemplo, se houve:

> 1. Agência e supervisão humana:
>
> 1.1. Avaliação de impacto sobre os direitos fundamentais, identificando e documentando possíveis trocas entre os diferentes princípios e direitos[60];
> 1.2. Avaliação da interação do sistema de IA com decisões de usuários humanos (se houve ações recomendadas, apresentação de opções ou decisões a serem tomadas pelo usuário)[61];
> 1.3. Avaliação sobre a possibilidade de o sistema de IA afetar a autonomia humana, interferindo no processo de tomada de decisão do usuário[62];
> 1.4. Avaliação da presença da comunicação aos usuários de que a decisão, o conteúdo, o conselho ou o resultado do sistema são resultado de uma operação algorítimica e de que estão interagindo com um agente não humano[63];
> 1.5. Avaliação da presença de medidas de supervisão e controle humanos apropriados, descrevendo o nível de controle ou envolvimento e as ferramentas de intervenção[64];
> 1.6. Avaliação da capacidade do sistema de aprimorar ou aumentar capacidades humanas[65];
> 1.7. Avaliação da presença de medidas de segurança que evitem excesso de confiança e tornem prudente a condução do procedimento[66];

---

[59] MORAES, Maria Celina Bodin de. A constitucionalização do direito civil e seus efeitos sobre a responsabilidade civil. *Direito, Estado e Sociedade,* Rio de Janeiro, n. 29, p. 233-258, 2006.
[60] "Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?" (p. 26).
[61] "Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?" (p. 26).
[62] "Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?" (p. 26).
[63] "In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent? Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision?" (p. 26).
[64] "Did you consider the appropriate level of human control for the particular AI system and use case? Did you put in place mechanisms and measures to ensure human control or oversight? Did you take any measures to enable audit and to remedy issues related to governing AI autonomy (p. 26).
[65] "Does the AI system enhance or augment human capabilities?"(p. 26).
[66] "Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?" (p. 26).

1.8. Avaliação de medidas que permitam à auditoria remediar questões relacionadas ao controle da autonomia da IA[67];

2. Robustez técnica e segurança:

2.1. Avaliação de possíveis formas de ataque e vulnerabilidades do sistema de IA e as respectivas medidas para garantir a integridade e a resiliência do sistema contra tais disfunções, descrevendo riscos e seguranças[68];
2.2. Avaliação do comportamento do sistema em situações e ambientes inesperados[69];
2.3. Avaliação de plano de recuperação na hipótese de ataques adversários ou situações inesperadas[70];
2.4. Avaliação da comunicação ao usuário dos riscos apresentados pelo sistema de IA e da existência de planos para mitigar ou gerenciar tais riscos[71];
2.5. Avaliação da presença de apólices de seguro para lidar com possíveis danos do sistema de IA[72];
2.6. Avaliação da probabilidade de o sistema de IA causar danos aos usuários ou a terceiros, bem como ao meio ambiente ou aos animais[73];

3. Privacidade e governança de dados

3.1. Avaliação da existência de mecanismos que permitam que outras pessoas assinalem problemas relacionados à privacidade ou proteção de dados nos processos de coleta e processamento[74];
3.2. Avaliação do tipo e escopo de dados armazenados (por exemplo, se há dados sensíveis ou pessoais)[75];
3.3. Avaliação da possibilidade de desenvolver o sistema de IA ou treinar o modelo sem ou com o uso mínimo de dados potencialmente confidenciais ou pessoais[76];
3.4. Avaliação da tomada de medidas que resguardem a privacidade, como criptografia ou anonimização[77];
3.5. Avaliação da presença de um *Data Protection Officer* (DPO)[78];

4. Transparência

[67] "Did you take any measures to enable audit and to remedy issues related to governing AI autonomy?" (p. 26).
[68] "Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks? Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?" (p. 26).
[69] "Did you verify how your system behaves in unexpected situations and environments?" (p. 26).
[70] "Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?" (p. 26).
[71] "Did you consider the level of risk raised by the AI system in this specific use case? Did you consider an insurance policy to deal with potential damage from the AI system? Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse? Is there a plan to mitigate or manage these risks?" (p. 26).
[72] "Did you consider an insurance policy to deal with potential damage from the AI system?" (p. 27).
[73] "Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity? Did you consider the liability and consumer protection rules, and take them into account?" (p. 27).
[74] "Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's processes of data collection (for training and operation) and data processing?" (p. 28).
[75] "Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?" (p. 28).
[76] "Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?" (p. 28).
[77] "Did you take measures to enhance privacy, such as via encryption, anonymisation and aggregation?" (p. 28).
[78] "Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?" (p. 28).

> 4.1. Avaliação da consideração no sistema de possíveis limitações humanas, tais como riscos de confusão, vieses prévios ou fadigas cognitivas[79];
>
> 5. Diversidade, não discriminação e justiça
>
> 5.1. Avaliação da acessibilidade do sistema aos usuários com deficiência, bem como a consulta de tal comunidade durante a fase de desenvolvimento do sistema[80];
> 5.2. Avaliação sobre a equipe envolvida na construção do sistema, verificando se é representativa do público-alvo e da população em geral, considerando também outros grupos que podem ser tangencialmente impactados[81];
> 5.3. Avaliação da presença de *feedback* de outras equipes ou grupos que representam diferentes origens e experiências[82];
>
> 6. Bem-estar social e ambiental
>
> 6.1. Avaliação do impacto ambiental do desenvolvimento, implantação e uso do sistema de IA[83];
> 6.2. Avaliação dos impactos sociais do sistema, tais como risco de perdas de empregos, bem como medidas que podem neutralizar tais riscos[84].
>
> 7. Prestação de contas
>
> 7.1. Avaliação do estabelecimento de mecanismos que facilitem a auditabilidade do sistema, como garantir a rastreabilidade e o registro de processos e resultados dos sistemas de IA[85];
> 7.2. Avaliação da presença de um conjunto adequado de mecanismos que permitem reparação em caso de ocorrência de algum dano ou impacto adverso[86];

O rol é exemplificativo e no texto das Diretrizes ainda existem outras sugestões de avaliações a serem feitas, que visam direcionar a atividade dos desenvolvedores de inteligência artificial, fazer com que a utilização do sistema seja cada vez mais acessível e democrática e diminuir os riscos apresentados pela tecnologia. É imprescindível que sejam estabelecidas balizas acerca da utilização da IA. Não é admissível que a tecnologia seja utilizada para o desenvolvimento de sistemas de pontuação e monitoramento de cidadãos[87], o que já tem sido cogitado e acarreta inegáveis discriminações e violações aos direitos fundamentais.

O desenvolvimento tecnológico deve ocorrer em consonância com a proteção da liberdade e autonomia humana, o que demanda esforços da doutrina jurídica na compreensão

---

[79] "Depending on the use case, did you think about human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue?" (p. 29).
[80] "Did you assess whether the AI system usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified?" (p. 29).
[81] "Did you assess whether the team involved in building the AI system is representative of your target user audience? Is it representative of the wider population, considering also of other groups who might tangentially be impacted?" (p. 30).
[82] "Did you get feedback from other teams or groups that represent different backgrounds and experiences?" (p. 30).
[83] "Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)?" (p. 30).
[84] "Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?" (p. 31).
[85] "Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?" (p. 31).
[86] Você estabeleceu um conjunto adequado de mecanismos que permitem reparação em caso de ocorrência de algum dano ou impacto adverso? (p. 31).
[87] Disponível em: https://www.bbc.com/portuguese/internacional-42033007. Acesso em: 24 maio 2019.

das capacidades e limitações desse fenômeno. Ao tempo que se deve assegurar a solidariedade social e a reparação integral dos danos, é preciso atentar para a determinação do art. 218 da Constituição Federal:

> o *caput* do art. 218 estabelece que é dever do Estado, União, Estados e Municípios promover e incentivar o desenvolvimento científico, a pesquisa e a capacitação tecnológicas. De modo que emerge do citado texto nítida separação entre, de uma banda o desenvolvimento científico, e de outra a pesquisa e capacitação tecnológicas, em consonância com dois objetivos fundamentais da CF/88, nos termos do art. 3º, II, assegurando o desenvolvimento nacional, e o disposto no art. 3º, III, erradicando a pobreza e a marginalização e reduzindo as desigualdades sociais e regionais, fundando-se, sobretudo, na soberania, dignidade da pessoa humana e nos valores sociais do trabalho e da livre iniciativa, art. 1º, incisos I, III e IV. Ressalte-se o enfoque claro no interesse e solução dos problemas brasileiros, assim como no efetivo desenvolvimento nacional e diminuição das diferentes e enormes desigualdades e contrastes sociais e espaciais, além da defesa intransigente da soberania política[88].

A problemática exsurge, assim, a partir da perspectiva de que a solução deve perpassar pela necessária compatibilização entre o princípio da reparação integral do dano da vítima e o incentivo ao desenvolvimento de inovação e novas tecnologias, estipulado no art. 218 da Constituição Federal, visando a um equilíbrio entre valores imprescindíveis no ordenamento jurídico.

Os desafios são múltiplos e a doutrina ainda não possui respostas para todos os questionamentos que decorrem do desenvolvimento tecnológico, especialmente considerando o intrínseco dinamismo que impera nesse âmbito. A tecnologia e sua relação no meio social são como um vírus: estão em constante mutação. É um fenômeno que não pede licença nem tem pretensão de ir embora, restando, por ora, o desafio de disciplinar essas questões.

6 CONSIDERAÇÕES FINAIS

Por que se propugna, afinal, pela sistematização de diretrizes éticas sobre a confiabilidade da inteligência artificial? Trata-se, com efeito, de agregar esforços na tentativa de controle social do desenvolvimento tecnológico, para que a automação das máquinas não ameace os direitos fundamentais e todas as garantias consolidadas no Estado Democrático de Direito.

Para tanto, é imprescindível que a autonomia humana, a prevenção de danos, a tutela dos vulneráveis e a explicabilidade do sistema sejam observadas. Além disso, é necessário que haja supervisão humana, robustez e segurança técnica, privacidade e governança de dados, transparência, diversidade, respeito ao bem-estar social e ambiental e responsabilização. Em termos concretos, significa dizer que avaliações de impactos precisam ser feitas, especialmente no que tange à interação do sistema com os usuários, medidas de supervisão, segurança, auditoria e comunicação de riscos.

---

[88] VEGA GARCIA, Balmes. *Direito e tecnologia*: regime jurídico da ciência, tecnologia e inovação. São Paulo: LTr, 2008. p.110.

Não se trata de tarefa fácil, máxime tendo em vista que o art. 218 da Constituição Federal demanda que toda regulação do setor tecnológico seja procedida da cautela necessária a não obstaculizar o seu desenvolvimento, sendo dever dos entes federativos promover e incentivar o desenvolvimento científico, a pesquisa e a capacitação tecnológicas.

O avanço da inteligência artificial no contexto social é inesgotável. Cumpre desenvolver formas de assegurar que os riscos e impactos adversos da tecnologia sejam tratados de maneira adequada. Nesse diapasão, as Diretrizes Europeias apontam que a IA possui três atributos primordiais: deve observar a legalidade, garantindo a conformidade com as leis vigentes; deve ser ética, assegurando a conformidade com os princípios e os valores axiológicos; deve ser robusta, tanto de uma perspectiva técnica quanto social, para que o sistema cause o mínimo possível de danos.

Não se ignora que a abertura semântica e a relativização de quais seriam os conceitos éticos aplicados no caso concreto podem vir a ensejar dificuldades, tornando abstratas as prescrições estipuladas no documento. Ademais, é possível que surjam tensões entre os próprios interesses jurídicos envolvidos. No entanto, tais diretrizes podem ser incorporadas no ordenamento jurídico pátrio, onde ainda há certo vácuo legislativo no que tange à regulação da inteligência artificial, servindo como ponto de partida para a disciplina da questão, especialmente no que concerne aos parâmetros principiológicos. Propugna-se, nesse sentido, pela consolidação de uma perspectiva tecnológica que assegure o respeito aos direitos fundamentais, à democracia e ao Estado de Direito.

REFERÊNCIAS

ALEXY, Robert. *Teoria dos direitos fundamentais*. Tradução de Virgílio Afonso da Silva. 2. ed. São Paulo: Malheiros, 2015.

CÂMARA, Marco Sérgio Andrade Leal Câmara. *Inteligência artificial*: representação de conhecimento. Disponível em: https://student.dei.uc.pt/~mcamara/artigos/inteligencia_artificial.pdf. Acesso em: 22 set. 2018.

CERKA, Paulius; GRIGIENE, Jurgita; SIRBIKYTE, Gintare. Liability for damages caused by artificial intelligence. *Computer Law and Security Review*, Londres, v. 31, n. 3, p. 376-389, jun. 2015.

CORREIA JÚNIOR, José Barros. *A função social e a responsabilidade social da empresa perante os stakeholders.* 2013. Tese (Doutorado em Direito). Pós-Graduação em Direito da Faculdade de Direito de Recife da Universidade Federal de Pernambuco, Recife.

CUNHA JÚNIOR, Dirley da. *Curso de Direito Constitucional.* 3. ed. rev. ampl. e atual., Salvador: Juspodivm, 2009.

DONEDA, Danilo Cesar Maganhoto; MENDES, Laura Schertel; SOUZA, Carlos Affonso Pereira de; ANDRADE, Norberto Nuno Gomes de. Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal. *Pensar: Revista* de Ciências Jurídicas, Fortaleza, v. 23, n. 4, p. 1-17, out./dez. 2018.

HALLEVY, Gabriel. The criminal liability of artificial intelligence entities- from Science fiction to legal social control. *Akron Intellectual Property Journal*, Akron, v. 4, p. 171-199, 2016.

HENDERSON, Harry. *Artificial intelligence*: mirrors for the mind. New York: Chelsea House Publishers, 2007.

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINE. *Ethics Guidelines for trustworthy AI.* Disponível em: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Acesso em: 4 maio 2019.

KAPLAN, Jerry. *Artificial Intelligence*: What everyone needs to know. Oxford: Oxford University Press, 2016.

KNIGHT, Will. *The dark secret at the heart of AI.* Disponível em: https://www.technology review.com/s/604087/the-dark-secret-at-the-heart-of-ai/ Acesso em: 26 set. 2018.

LÔBO, Paulo. A informação como direito fundamental do consumidor. *Revista de Direito do Consumidor*, São Paulo: Revista dos Tribunais, ano 10, n. 37, 2001.

MORAES, Maria Celina Bodin de. A constitucionalização do direito civil e seus efeitos sobre a responsabilidade civil. *Direito, Estado e Sociedade,* Rio de Janeiro, n. 29, p. 233-258, 2006.

NORVIG, Peter; RUSSELL, Stuart J. *Artificial Intelligence*: A Modern Approach. New Jersey: Prentice Hall, 1995.

OLIVEIRA, Samuel Rodrigues De; COSTA, Ramon Silva. Pode a máquina julgar? Considerações sobre o uso de inteligência artificial no processo de decisão judicial. *Revista de Argumentação e Hermenêutica Jurídica*, Porto Alegre, v. 4, n. 2, p. 21-39, jul./dez 2018.

PEIXOTO, Erick Lucena Campos; EHRHARDT JÚNIOR, Marcos. Breves notas sobre a ressignificação da privacidade. *Revista Brasileira de Direito Civil – RBDCivil*, Belo Horizonte, v. 16, p. 35-56, abr./jun. 2018.

PORTO, Uly de Carvalho Rocha. *A responsabilidade civil extracontratual por danos causados por robôs autônomos.* 2018. Dissertação (Mestrado em Ciências Jurídico-Civilistas), Faculdade de Direito da Universidade de Coimbra, Coimbra, 128 p.

RODOTÀ, Stefano. *A vida na sociedade de vigilância*: a privacidade hoje. Tradução de Danilo Doneda e Luciana Cabral Doneda. Rio de Janeiro: Renovar, 2008.

STRELKOVA, O. PASICHNYK, O. *Three types of artificial intelligence.* Disponível em: http://eztuir.ztu.edu.ua/jspui/bitstream/123456789/6479/1/142.pdf. Acesso em: 3 maio 2019.

TOMASEVICIUS FILHO, Eduardo. Inteligência artificial e direitos da personalidade: uma contradição em termos? *Revista da Faculdade de Direito da Universidade de São Paulo*, São Paulo, v. 113, p. 133-149, jan./dez. 2018.

UNIÃO EUROPEIA. *Carta dos Direitos Fundamentais da União Europeia.* Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:12016P/TXT&from=EN. Acesso em: 24 set. 2019.

VEGA GARCIA, Balmes. *Direito e tecnologia*: regime jurídico da ciência, tecnologia e inovação. São Paulo: LTr, 2008.

VIGNE, Vernor. *What is the singularity?* Disponível em: https://www.frc.ri.cmu.edu/~hpm/ book98/com.ch1/vinge.singularity.html Acesso em: 26 set. 2018.

WINSTON, Patrick Henry. *Artificial Intelligence.* 3. ed. Boston: Addison-Wesley Publishing Company, 1993.

Recebido: 25.07.2020

Aprovado: 29.09.2020

Como citar: SILVA, Gabriela Buarque Pereira; EHRHARDT JÚNIOR, Marcos. Diretrizes éticas para a Inteligência Artificial confiável na União Europeia e a regulação jurídica no Brasil. **Revista IBERC**, Belo Horizonte, v. 3, n. 3, p. 1-28, set./dez. 2020.