RISCOS E EXPLICABILIDADE A PARTIR DA INTELIGÊNCIA ARTIFICIAL COMO NÃO-COISA

RISKS AND EXPLICABILITY FROM ARTIFICIAL INTELLIGENCE AS NON-THING

Cinthia Obladen de Almendra Freitas[1]

Artigo Convidado

Resumo

Partindo-se da premissa de que a IA é composta por algoritmos que atuam sobre dados e adotando-se o método dedutivo, assume-se a IA como não-coisa e discute-se sobre riscos decorrentes de sistemas de IA, trabalhando especialmente os riscos de não se entender um algoritmo e a tomada de decisão automatizada. Assim, apresenta-se e trabalha-se a necessidade de se alcançar a explicabilidade em sistemas de IA, ou IA Explicável ou *AI Explainable* ou XAI, passando por responsabilidade, complexidade, verificabilidade e transparência. O artigo possui uma concepção jurídico-filosófica a partir da filosofia de Byung-Chul Han (2022) sobre não-coisas, trabalhando os riscos advindos de sistemas de IA e a explicabilidade de tais sistemas como elementos que permeiam o futuro da aplicação dos sistemas de IA na Era das Não-Coisas.

Palavras-chave
Direito e tecnologia; Inteligência artificial; Não-coisas; Riscos; IA explicável.

Abstract

Based on the premise that AI is composed of algorithms that act on data and adopting the deductive method, AI is assumed to be a non-thing and the risks arising from AI systems are discussed, especially the risks of not understanding an algorithm and automated decision-making. Thus, the need to achieve explainability in AI systems, or Explainable AI or XAI, is presented and addressed, encompassing responsibility, complexity, verifiability, and transparency. The article has a legal-philosophical conception based on Byung-Chul Han's (2022) philosophy on non-things, addressing the risks arising from AI systems and the explainability of such systems as elements that permeate the future of the application of AI systems in the Age of Non-Things.

Keywords
Law and technology; Artificial intelligence; Non-things; Risks; *AI explainable.*

[1] Doutora em Informática pela Pontifícia Universidade Católica do Paraná, Brasil, Professora Titular da Escola de Direito e do Programa de Pós-Graduação em Direito da mesma instituição, cinthia.freitas@pucpr.br.

FREITAS, Cinthia Obladen de Almendra. Riscos e explicabilidade a partir da inteligência artificial como não-coisa. **Revista Democracia Digital e Governo Eletrônico**, Florianópolis, v. 1, n. 24, p. 31-55, 2025. Seção A. Edição Especial do 33º Encontro Ibero Americano de Governo Eletrônico e Inclusão Digital.

1 Introdução

Um objeto de pesquisa é definido a partir da observação da realidade. E, a realidade atual exige um olhar diferenciado e ao mesmo tempo de incerteza, encantamento e necessária compreensão sobre a Inteligência Artificial. Freitas (2024) mostrou a necessidade de compreensão da IA tanto do ponto de vista tecnológico quanto jurídico, uma vez que o paradigma atual tem por base o Direito das Coisas, arraigado na posse e na coisa física, material e tangível, precisando urgentemente de uma releitura. E tal releitura encontra suporte na filosofia de Byung-Chul Han (2022) sobre não-coisas.

Inicialmente, tem-se por premissa que a Inteligência Artificial depende de 02 (dois) elementos que são a base do seu funcionamento: (i) algoritmos e (ii) dados. E, buscando compreender e definir IA, há que se pontuar que a IA é um ramo da Ciência da Computação e sua definição é bastante ampla e controversa.

Pode-se definir IA de diferentes maneiras, porém um ponto de partida é ter como premissa os tipos de problemas que a IA pode resolver. A partir do exposto, pode-se mencionar também que qualquer definição de IA será também complexa, visto envolver um vasto conjunto de tecnologias que se desenvolvem cada vez mais rápido e que podem trazer benefícios econômicos, sociais e ambientais. São alguns exemplos de benefícios: melhorar atividades de previsão, otimizar as operações e a aplicação de recursos, personalizar soluções digitais disponíveis para indivíduos e organizações, trazer vantagens competitivas às empresas e apoiar resultados social e ambientalmente benéficos, em áreas tais como: saúde, agricultura, educação e formação, gestão de infraestruturas, energia, transportes e logística, serviços públicos, segurança, justiça, eficiência energética e de recursos e, ainda, mitigação de riscos sociais (vulnerabilidade e exclusão digital) e ambientais (alterações climáticas).

Assim, do ponto de vista tecnológico, a IA pode ser descrita como o uso da tecnologia para automatizar tarefas que normalmente requerem inteligência humana, sendo "*The capacity of computers or other machines to exhibit or simulate inteligente behaviour*"[1] (OXFORD LIVING DICTIONARIES, 2023). De acordo com a Agência dos Direitos Fundamentais da União Europeia (2021, p. 01), a partir do documento denominado "Preparar o Futuro – Inteligência Artificial e Direitos Fundamentais – Síntese" tem-se que "Não existe uma definição de IA universalmente aceita. Ao invés de se referir a aplicações concretas, a mesma reflete a recente evolução tecnológica que engloba diversas tecnologias." O referido documento adota o conceito de IA definido em termos gerais, remetendo ao trabalho realizado pelo Grupo de Peritos de Alto Nível em Inteligência Artificial da Comissão Europeia (HLEG, 2019, p. 01), de modo que:

> O conceito de inteligência artificial (IA) aplica-se a sistemas que apresentam um comportamento inteligente, analisando o seu ambiente e tomando medidas — com um determinado nível de autonomia — para atingir objetivos específicos. Os sistemas baseados em inteligência artificial podem estar puramente confinados ao "software", atuando no mundo virtual (por exemplo: assistentes de voz, programas de análise de imagens, motores de busca, sistemas de reconhecimento facial e de discurso), ou podem estar integrados em dispositivos físicos (por exemplo: robôs avançados, automóveis autónomos, veículos aéreos não tripulados ou aplicações da Internet das coisas).

O HLEG observa e discute o fato do termo IA conter uma referência explícita à noção de inteligência, sendo que tal discussão pode ser ampliada para inteligência e cérebro ou, ainda, inteligência e mente, permitindo discussões complexas sobre homem e máquina, sua relação e como não somente criar ou imputar inteligências em máquinas, mas como reconhecer que máquinas são ou serão inteligentes. Há em tudo isso uma propriedade bastante importante a ser discutida, ou seja, a racionalidade: "Refere-se à capacidade de escolher a melhor ação a tomar para atingir um determinado objetivo, dados determinados critérios a otimizar e os recursos disponíveis." (HLEG, 2019, p. 01).

Importante observar que o HLEG (2019) aplica o termo sistema de IA ou baseado em IA referindo-se a qualquer componente que tenha em sua formação o uso de métodos e técnicas de IA. Entende-se, portanto, que sistemas de IA envolvem em seu *core* programas de computador (*software*), não eliminando a possibilidade se estarem integrados em máquinas (*hardware*), mas não é o processamento propriamente dito (realizado em máquinas, *hardware*) que define um sistema de IA, mas a escolha dos métodos e técnicas a serem aplicados na solução de um problema específico.

Por isso, IA é essencialmente *software*, ou seja, a conjunção de algoritmos que operam sobre dados, usando regras simbólicas ou treinando um modelo numérico (matemático, estatístico ou probabilístico), e também adaptando o seu funcionamento (comportamento) a partir da análise de como o ambiente (físico ou digital) é afetado pelos seus resultados anteriores. Por esses motivos, o HLEG (2019, p. 03-04) apresenta a IA como uma disciplina científica.

Por isso construir uma definição para IA ou mesmo sistemas de IA não é trivial. Mas é necessário ter olhos para o primeiro regramento de IA, denominado *AI Act*, desenvolvido pelo *European Parliamentary Research Service* (EPRS) da Comissão Europeia, no qual encontra-se a definição de IA em seu artigo 3(1), também considerando IA como sistema de IA (UNIÃO EUROPEIA, 2024, p. 112): "*is a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments*".

Entende-se, portanto, que um sistema de IA é *software* implementado e executado em *hardware*, que tem por base algoritmos e dados, projetado para operar com diferentes níveis de autonomia, visando atender objetivos explícitos ou implícitos de forma a gerar resultados, a exemplo de: previsões, recomendações, ou decisões que possam influenciar ambientes físicos ou virtuais.

Outras duas definições são relevantes para o estudo, a saber: (i) ponto de vista jurídico: Projeto de Lei 2338/2023 e (ii) ponto de vista tecnológico: norma técnica ABNT NBR ISO/IEC 22989 de 2023. A iniciativa de legislação brasileira para regular a IA parte da seguinte definição, conforme artigo 4º., inciso I (BRASIL, 2024, p. 19):

> I - sistema de inteligência artificial (IA): sistema baseado em máquina que, com graus diferentes de autonomia e para objetivos explícitos ou implícitos, infere, a partir de um conjunto de dados ou informações que recebe, como gerar resultados, em especial, previsão, conteúdo, recomendação ou decisão que possa influenciar o ambiente virtual, físico ou real;

Por outro lado, a definição apresentada na norma técnica NBR ISO/IEC 22989:2023 considera (ABNT, 2023, p. 02):

> É um sistema desenvolvido que gera saídas como conteúdo, previsões, recomendações ou decisões para um determinado conjunto de objetivos definidos pelo homem. O sistema desenvolvido pode utilizar diversas técnicas e abordagens relacionadas à inteligência artificial para desenvolver um modelo, para representar dados, conhecimento, processos etc. que podem ser usados para realizar tarefas.

O exercício de definir Inteligência Artificial, especificamente, sistemas de IA, foi relevante para o estudo realizado visto fornecer uma direção a ser seguida na compreensão dos demais aspectos relativos aos riscos e à explicabilidade de sistemas de IA.

Para realização da pesquisa foi aplicado o método dedutivo. Partindo-se da premissa de que a IA é composta por algoritmos que atuam sobre dados, assume-se a IA como não-coisa e discute-se sobre riscos decorrentes de sistemas de IA, trabalhando especialmente os riscos de não se entender um algoritmo e a tomada de decisão automatizada. Assim, apresenta-se e trabalha-se a necessidade de se alcançar a explicabilidade em sistemas de IA, ou IA Explicável ou *AI Explainable* ou XAI, passando por responsabilidade, complexidade, verificabilidade e transparência. O artigo possui uma concepção jurídico-filosófica a partir da filosofia de Byung-Chul Han (2022) sobre não-coisas, trabalhando os riscos advindos de sistemas de IA e a explicabilidade de tais sistemas como elementos que permeiam o futuro da aplicação dos sistemas de IA na Era das Não-Coisas. O artigo é resultado de projeto de pesquisa aprovado na Chamada Universal 10/2023 do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

2 A Inteligência artificial como não-coisa

A reflexão sobre a Inteligência Artificial como não-coisa advém da filosofia de Byung-Chul Han (2022) e esse tema foi explorado por Freitas (2024) ao buscar aspectos tecnológicos e jurídicos da IA. No que se refere aos aspectos tecnológicos, Freitas (2024, p. 89-99) explica sobre dados e algoritmos, passando pelos aspectos fundamentais da IA. Já nos aspectos jurídicos, Freitas (2024, p. 99-105) retoma o Direito das Coisas, o qual enfrenta uma nova realidade, visto que as coisas não são físicas, materiais e tangíveis no mundo digital. Para a autora "Tem-se aqui um ponto de inflexão, uma necessidade de mudança radicalmente oposta, porém a teoria precisa ser complementar, visto não serem antagônicas, mas trechos diferentes de uma mesma trajetória comum à ciência do Direito." (FREITAS, 2024, p. 100).

E, neste sentido, Freitas (2024) propõe uma análise da IA como não-coisa a partir de Byung-Chul Han (2022), relacionando o estudo da coisa e da não-coisa, agora centrada na informação obtida a partir do processamento de dados digitais e nos atores que processam dados e informações, vigiando e controlando cada um dos seres humanos da infoesfera (FLORIDI, 2014).

A proposta de Freitas (2024) é possível uma vez que a IA é composta basicamente por dados e algoritmos, os quais são composta por conjuntos de bits (*binary digit*), de zeros (0) e uns (1). O digital refere-se à representação por meio de valores discretos (0 e 1), portanto, uma

representação matemática do descontínuo, mas quando corretamente composta forma a representação de um todo, de uma coisa para o mundo físico, analógico. Por isso, Freitas (2024, p. 100) contextualiza sua análise a partir de 04 (quatro) propriedades físicas que fundamentam o olhar jurídicos das coisas considerando-se uma leitura a partir de aspectos tecnológicos do mundo digital, sendo tais propriedades: (i) materialidade, (ii) pessoalidade, (iii) territorialidade, (iv) temporalidade e (v) espacialidade.

E, toda essa análise nasce do entendimento de Byung-Chul Han sobre não-coisas. Han (2022, p. 9) preconize que a "ordem terrena está a ser substituída pela ordem digital. Esta desreifica o mundo, ao mesmo tempo que o informatiza." Para o autor, as não-coisas "penetram de todos os lados no meio que nos rodeia e ocupam o lugar das coisas". De modo que "nos encontramos na transição da era das coisas para a era das não-coisas." E as não-coisas se tornaram sinônimo dos dados e informações que determinam a vida no mundo digital ou a *onlife* na infosfera (FLORIDI, 2014, p. 40-41).

Yuval Noah Harari (2016, p. 370-399), em seu livro "Homo Deus: uma breve história do amanhã", discute a religião dos dados, ou dataísmo (tradução para o português de *dataism*, sendo o termo *data* mantido em inglês para o termo que se refere aos dados), pelo qual "o Universo consiste num fluxo de dados e o valor de qualquer fenômeno ou entidade é determinado por uma contribuição ao processamento de dados." E, o autor aponta que "o supremo valor dessa religião é o fluxo de informação". E, Harari, ousa deixar os leitores com 03 (três) questionamentos: (i) Será que organismos são apenas algoritmos e a vida apenas processamento de dados?; (ii) O que é mais valioso – inteligência ou consciência?; (iii) O que vai acontecer à sociedade, aos políticos e à vida cotidiana quando algoritmos não conscientes mas altamente inteligentes nos conhecerem melhor do que nós nos conhecemos? São muitas as perguntas que precisam de respostas.

Todo esse cenário conta ainda com os estudos de Shoshana Zuboff (2021) sobre o que ela nomeia como capitalismo de vigilância, tendo como base uma arquitetura global de modificação comportamental que altera e desfigura o mundo físico, visto que a nova configuração global tem por fundamento uma estruturação digital por meio dos dados que os usuários "deixam" na Internet e são tratados sem o devido consentimento. A autora exemplifica e discute fortemente o tratamento de dados frente a ausência ou fraco regramento de proteção de dados pessoais e ao desconhecimento tanto da coleta e tratamento de dados quanto dos reflexos na ordem social e futuro digital da sociedade contemporânea.

E Han (2022, p. 45-50) estabelece um paralelo entre o pensamento humano e a Inteligência Artificial para apresentar a IA como não-coisa. Tal paralelo parte da premissa que o pensamento é composto por uma "totalidade, que precede os conceitos, as ideias e a informação", de modo a se encontrar "numa disposição afetiva básica" (HAN, 2022, p. 45), a qual se coloca para fora do ser humano. Ou seja, o pensamento pode ser externado, para fora de si. Já a IA, ao realizar cálculos, "nunca está fora de si mesma" (HAN, 2022, p. 46). Os cálculos não envolvem emoção e estar fora de si mesma é emoção. Falta à IA o espírito. Portanto, entende-se que IA é existência em bits, mas não é essência.

Para Han (2022, p. 47) "o pensamento ouve, ou melhor, escuta e presta toda a atenção. A Inteligência Artificial é surda." Tudo isso compõe a dimensão analógica do pensamento humano, a qual não se pode reproduzir por meio da IA (HAN, 2022, p. 47). E, Han (2022, p. 48) corrobora essa explicação, afirmando que "A Inteligência Artificial é apática, ..., sem paixão. Só calcula." Para os seres humanos, existem fatos, contextos e mudanças. Para a IA existem dados, que podem ser revistos e modificados para efetuar novos cálculos e obter novos resultados, até mesmo retroalimentar os dados já existentes em uma base de dados. Mas tudo isso depende de um algoritmo que precisará ser planejado para tal tarefa. O ser humano não necessita de algoritmo, ele mesmo lhe dá novos fatos e contextos, gerando mudanças e entendimentos. Han (2022, p. 49) explica que o contexto é compreendido pelo ser humano, por meio do pensamento. Para a IA, a relação entre A e B, expressa por C, somente é compreendida se o algoritmo assim estiver previsto para relacionar A e B e encontrar C. Por exemplo, se A representa o valor 10 e B representa o valor 20, C poderá ser positivo ou negativo a depender da relação matemática que se deseja analisar, ou seja, se A é maior, menor ou igual a B. Portanto, o contexto para os seres humanos é relacional e para a IA é um o resultado de uma análise matemática lógica.

Outra característica dos seres humanos e a capacidade, por meio da inteligência, de realizar escolhas, sendo que a IA "só opta por uma escolha entre opções previamente dadas", sendo o resultado positivo (1) ou negativo (0) ou, ainda, uma probabilidade (um valor numérico entre 0 e 1 ou 0 e 100%). O pensamento é não determinístico e infinito, enquanto um algoritmo que opera com dados é determinístico e finito. Assim IA não é baseada em pensamentos, mas em algoritmos que representam métodos e técnicas de diferentes ramos de aplicação, por exemplo, visão computacional (*computer vision*), processamento de linguagem natural (*natural language processing*) ou robótica (*robotics*).

De um modo geral, os métodos e técnicas baseados em IA, funcionam tendo como *input* uma base de dados que reflete situações e cenários do passado, de modo que, por exemplo, tal qual como explicado por Freitas e Barddal (2019, p. 110) "a análise preditiva é uma abordagem popular para obter informações e padrões sobre os dados e criar modelos preditivos. A análise preditiva visa aproveitar os dados do passado para obter informações em tempo real e prever eventos futuros". Continuam os autores explicando que "Na prática, a análise preditiva está na interseção entre a estatística, matemática e ciência da computação, que, em sua influência, pode ser aplicada para obter insights e ganhos em diferentes aplicações." Han (2022, p. 50) contradiz a questão da IA prever eventos futuros, visto que "o futuro que calcula não é um futuro no sentido próprio do termo." Entende-se que para Han o futuro calculado pela IA não tem contexto, semântica ou significado, sendo somente uma representação numérica (matemática, lógica, estatística ou probabilística). O contexto, a semântica ou o significado serão postos pelos seres humanos.

Por isso, Han (2022, p. 51) afirma que "A informação e os dados não tem profundidade", visto que contexto, semântica ou significado surgem do pensamento humano que "é mais do que cálculos e resolução de problemas". "Os dados e a informação não seduzem" (HAN, 2022, p. 51). O que seduz é o pensamento do ser humano ao envolver-se com o resultado (*output*) fornecido pela IA. Não seduzem, uma vez que para a IA são dados, *input* para os algoritmos, portanto, representação matemática em bits de coisas. Só o pensamento vê,

escuta ou entende a coisa representada. A coisa torna-se não-coisa por meio dos dados e algoritmos. E foi a não-coisa que passou a influenciar a sociedade contemporânea e estabeleceu a sociedade de algoritmos (SHENK, 1997; BRIN, 1998; SCHUILENBURG; PEETERS, 2021) na era das não-coisas de Byung-Chul Han.

3 Riscos e sistemas de IA

A realidade vem demonstrando que a regulação dos sistemas de IA partem de abordagens baseadas em riscos (*risk-based regulatory approach*). E estudar riscos é igualmente importante por sistemas de IA envolverem o tratamento de dados, entre estes, dados pessoais e sensíveis e estarem intrinsicamente ligados tanto aos direitos fundamentais quanto à complexidade e necessidade de explicabilidade. E, considerando sistemas de IA como não-coisa, todos estes elementos precisam ser estudados e justapostos para que as relações adentrem um mesmo espaço epistemológico para se alcançar os fundamentos lógicos, tecnológicos e jurídicos, o valor propriamente dito dos sistemas de IA na sociedade contemporânea e de algoritmos, sem esquecer da importância objetiva dos sistemas de IA como não-coisa.

O estudo inicia-se com o entendimento do que é risco e cabe destacar que risco não é sinônimo de perigo, como discutido por Freitas (2023), de modo a ampliar e atualizar o estudo que fora realizado. O trabalho de Raphaël Gellert (2017, p. 02) explica que existem 02 (dois) significados para risco: um vernáculo e outro técnico. Gellert se apoia em outro autor, Godard et al. (2002, p. 12) e assim explicam: (i) no sentido vernáculo, o risco é geralmente referido como futuro, possível perigo, ou seja, como "um eventual perigo que só pode ser previsto até certo ponto"[2] e (ii) no sentido técnico, risco é usado para a tomada de decisões com base em avaliação de eventos futuros, sendo que seus elementos constitutivos são compostos por duas operações distintas, mas unidas e dependentes entre si, quais sejam: a) prever eventos futuros (tanto negativos quanto positivos) e tomar decisões com base nesses eventos.

Cabem aqui 02 (dois) contrapontos. O primeiro explicita que há que se refletir sobre os riscos decorrentes da inter-relação e interdependência do homem contemporâneo com o meio ambiente digital, conceito estabelecido em Cavedon et al. (2015). O segundo contraponto é que deve-se ter por base que o tratamento de dados pessoais está (ou pode estar) vinculado ao surgimento de riscos capazes de comprometer a qualidade de vida do homem (titulares de dados), considerando-se o titular de dados como parte indissociável do meio em que vive e com o qual, necessariamente, interage. Especialmente, se o meio for o meio ambiente digital.

Assim, Cavedon et al. (2015, p. 197) explicam a partir de Pardo (1999, p 25-26) que, "enquanto os perigos têm causas essencialmente naturais, os riscos surgem a partir do momento em que os seres humanos passam a interferir no curso da natureza". E no meio ambiente digital não é diferente. O tratamento de dados pessoais gera interferências na privacidade dos indivíduos e, portanto, na personalidade, ou seja, no conjunto de características que definem uma pessoa, seu padrão de individualidade pessoal e social.

[2] Texto original: "an eventual danger that can be foreseen only to some extent".

Corresponde a dizer que, segundo os autores, os perigos advêm a partir das variações próprias do ambiente, enquanto os riscos surgem da intervenção do homem no intuito de eliminar esses perigos, ou seja, o homem passa a interferir no meio em que vive e, como consequência, os riscos se manifestam. Nesse contexto, pode-se afirmar que a origem dos riscos se vincula diretamente aos processos de tomada de decisão, refletindo claramente o anseio humano de subjugar a natureza, sendo que tratar dados pessoais sem consentimento, como definido em diversos regramentos e legislações nacional ou internacionalmente, é subjugar os direitos dos titulares, por exemplo a partir de violações, usos indevidos, cookies, entre outros.

Pode-se, portanto, argumentar que qualquer decisão relativa ao risco envolve 02 (dois) elementos distintos e inseparáveis: "os fatos objetivos e uma visão subjetiva sobre a conveniência do que se ganha ou se perde com a decisão"[3] (BERNSTEIN, 1996, p. 100). E a tomada de decisão sobre o tratamento de dados pessoais não pode ser uma relação de perda para os titulares de dados.

Do ponto de vista técnico do significado de risco, deve-se ter em mente que risco é "combinação da probabilidade de um evento e de suas consequências", de acordo com ABNT ISO/IEC Guia 73 (2005). Tecnicamente, o risco é igual ao resultado da multiplicação da Consequência (C_i) pela Probabilidade (P_i); sendo que Consequência é o impacto ambiental caso ocorra um evento e Probabilidade é a probabilidade de ocorrência de um impacto que afete o meio ambiente, seja esse natural ou artificial no caso do meio ambiente digital. Outra definição de risco considera que "é um cenário que descreve um evento e suas consequências, estimado em termos de gravidade e probabilidade"[4] (DATA PROTECTION WORKING PARTY, 2017, p. 06). E, ainda, de acordo com a ISO/IEC 27001 (2022) risco é "o efeito da incerteza sobre os resultados desejados"[5], ou seja, há consequência(s) decorrente(s) do risco que estão sob um regime de incerteza (probabilidade de ocorrência).

A partir de Ulrich Beck (1998, p. 64) tem-se o entendimento de uma sociedade de risco, a qual tem sua origem quando as ameaças oriundas de ações e decisões humanas rompem os pilares de certeza estabelecidos pela sociedade industrial, minando, como consequência, os seus padrões de segurança. Sendo esse entendimento cabível na sociedade informacional em que se vive. No que tange aos dados pessoais, seja desde a coleta até seu descarte, passando por inúmeras operações de tratamento de dados, há que se considerar que os riscos abstratos, "além de imprevisíveis e incontroláveis, são também transfronteiriços e transtemporais" (CAVEDON et al., 2015, p. 200).

Transfronteiriços porque ultrapassam os limites do local originalmente impactado, por exemplo onde ocorre a coleta de dados pessoais. Isto posto, porque na visão de Beck (2004, p. 109-110) confronta-se o conceito de fronteira, entendido como parte limítrofe de um espaço em relação a outro. No meio ambiente digital, delimitar fronteiras é por vezes impossível. Dados fluem.

[3] Texto original: "the objective facts and a subjective view about the desirability of what is to be gained, or lose, by the decision".
[4] Texto original: "a scenario describing an event and its consequences, estimated in terms of severity and likelihood".
[5] Texto original: "the effect of uncertainty upon desired results".

E, são transtemporais porque não necessariamente se materializarão no momento em que se constituem, ou seja, a criação de um risco não implica, necessariamente, um dano imediato. Esse cenário é bem cabível em tratamento de dados pessoais, visto que dados coletados podem ser desviados para outra finalidade, sem que o titular consinta ou tenha conhecimento.

Diante dessas características, entende-se que Gellert (2017, p. 02) considera que o risco continua sendo uma noção abstrata que necessita de metodologias, modelos e processos que o implementem concretamente. E, a conjunção de aspectos técnicos e jurídicos é que irão permitir que se evite, previna, avalie, quantifique e, finalmente, se mitigue riscos, de modo a proteger dados pessoais.

Portanto, ao se ter por premissa que é impossível reduzir riscos à zero, independentemente da área de aplicação, há que se mitigar os riscos jurídicos e tecnológicos decorrentes do meio ambiente digital para se alcançar a proteção de dados pessoais e garantir direitos e liberdades dos titulares de dados.

Essa é a função da análise de risco, também chamada às vezes de gerenciamento de risco. Espelhando a dimensão dupla do risco, a análise de risco é composta de duas etapas (GELLERT, 2017, p. 02): (i) avaliação de risco: medir o nível de risco em termos de probabilidade e gravidade; (b) gerenciamento de risco: decidir se deve ou não assumir o risco. E, a decisão no nível de gerenciamento de risco é geralmente acompanhada de medidas que visam reduzir o nível de risco (GELLERT, 2017, p. 02). E existem muitos riscos jurídicos e tecnológicos relacionados ao tratamento de dados pessoais. Gellert (2017, p. 02) alerta que as medidas aplicadas na redução de riscos podem ser referidas por vezes como: redução de risco, controle de risco, resposta a risco ou, mais genericamente, como medidas de mitigação de risco, termo comumente utilizado na área jurídica.

E estar em conformidade com legislações ou regramentos, seja de proteção de dados pessoais (UNIÃO EUROPEIA, 2016; BRASIL, 2018) ou sobre sistemas de IA (UNIÃO EUROPEIA, 2024; BRASIL, 2024), é mitigar riscos. Mas há um ponto crucial nessa discussão que é determinar se o nível de risco é suficientemente baixo para que possa ser tomado. E a categorização do risco é muito importante quando o tema é Inteligência Artificial, especificamente, sistemas de IA.

Por isso, adentra-se ao tema de avaliação de risco, uma vez que alguns questionamentos precisam ser evidenciados, tais como: O que deve ser avaliado? O provável alto risco para os direitos e liberdades dos usuários-consumidores-cidadãos ou o impacto a partir da aplicação de sistemas de IA? São muitas perguntas e espera-se contribuir com essa discussão.

Inicialmente cabe exemplificar alguns riscos relacionados ao uso de sistemas de IA, conforme o *AI Act* e o PL N° 2.338/2024, frente aos riscos à vida ou à integridade física de indivíduos, sem que seja discutido ou analisado a categorização do risco contida nestes instrumentos. O Quadro 8 apresenta 04 (quatro) diferentes riscos envolvendo sistemas de IA, de modo que o cenário de ameaça considerado gera riscos e necessita compor uma análise de riscos por afetarem direitos e liberdades individuais. A relação de risco se estabelece por meio de

vulnerabilidades exemplificadas e cabíveis ao exemplo simulado. Os exemplos permitem compreender como um cenário de ameaça está relacionado com vulnerabilidades e riscos.

O Quadro 1 esclarece que os sistemas de IA estão sujeitos a riscos por meio de vulnerabilidades de cunho tecnológico, a partir de ameaças e vulnerabilidades de natureza disruptiva (CALDAS; FREIRE, 2013, p. 02). O exemplo torna claro que os sistemas de IA apresentam riscos aos direitos e liberdades dos usuários-consumidores-cidadãos, uma vez que impactam ora a tomada de decisão, ora a proteção de dados pessoais ou, ainda, podendo causar danos. Nota-se, portanto, que a análise de riscos tecnológicos não é trivial, especialmente quando relacionados à aplicação de sistemas de IA, havendo a necessidade de se distinguir entre análise de riscos e gestão de riscos.

Quadro 1: Exemplos de riscos relacionados à sistemas de IA.

Cenário de ameaça	**Vulnerabilidades**	**Riscos**
Uso de técnicas de *machine learning* para agrupar consumidores por categorias	Base de dados Número crescente de fontes primárias de coleta de dados pessoais (câmeras, sensores biométricos, geolocalização)	-Indução de comportamento (consumo de produtos e serviços) de forma prejudicial ou perigosa à saúde ou segurança ou contra fundamentos da Lei -Coleta de dados pessoais não necessários para fins primários -Divulgação de dados sensíveis sobre a vida de alguém -Recrutamento, triagem, filtragem ou avaliação discriminatória de candidatos/as

Fonte: A Autora.

Gellert (2017, p. 03) explica que a análise de risco é composta por etapas, a saber: (i) critérios de risco, (ii) identificação dos riscos e (iii) a avaliação de risco propriamente dita (ISO, 2009), sumarizadas a seguir:

- critérios de risco: compreende a definição de critérios para determinar se um evento pode ser considerado um risco, "os termos de referência contra os quais a significância de um risco é avaliada"[6] (ISO, 2009, p. 5). Parte importante dessa definição está nos procedimentos para identificar o que se apresenta como risco e como medir o nível de risco;
- identificação de risco: definida como o "processo de localização, reconhecimento e descrição de riscos"[7] (ISO, 2009, p. 5). Será necessário estabelecer uma comparação entre o evento em questão e os critérios estabelecidos no item anterior. O objetivo é determinar se o evento é suficientemente arriscado para ser considerado um risco;
- avaliação de risco: aplicação de metodologia para obtenção, por exemplo, de uma matriz de riscos. Após essa etapa é que os riscos poderão ser gerenciados em termos de custos e benefícios. Eis aqui a etapa de tomada de decisão para estabelecer os riscos que necessitam de ação mais urgente ou podem ser deixados em segundo plano.

[6] Texto original: "terms of reference against which the significance of a risk is evaluated".
[7] Texto original: "process of finding, recognizing and describing risks".

A necessidade de realizar a análise e a gestão dos riscos leva a compreensão de que a conformidade por si só, seja qual for a regulação em questão, não garante a capacidade de uma organização proteger dados pessoais. É necessário criar um vínculo robusto entre requisitos, políticas, objetivos, desempenho e ações voltadas à mitigação dos riscos.

A criação desse vínculo exige estabelecer os elementos constitutivos do risco, a saber (GELLERT, 2017, p. 03):

- 1º elemento: é o evento, o qual é definido pela ISO (2009, p. 4) como uma "ocorrência ou mudança de um determinado conjunto de circunstâncias"[8]. O evento pode ou não acontecer e terá uma série de consequências positivas e negativas na proteção de dados pessoais;
- 2º elemento: são as consequências, que são precisamente o "resultado de um evento"[9] (ISO, 2009, p. 5), quando tais impactos são negativos podem ser referidos como danos, e quando são positivos podem ser referidos como benefícios;
- 3º. terceiro: são os fatores de risco. Eles determinam se e como o risco se materializará, ou seja, é a probabilidade de ocorrência, bem como a sua gravidade. Sendo "elementos que, isoladamente ou em combinação, têm potencial intrínseco para gerar risco"[10] (ISO, 2009, p. 4).

Há, portanto, que se conhecer a probabilidade e a gravidade de um evento, definidos pelo *French Data Protection Authority* (CNIL, 2015) e explicado por Gellert (2017, p. 02-03;06-08): a) gravidade: representa a magnitude de um risco, dependendo principalmente da natureza prejudicial dos impactos potenciais e b) probabilidade: representa a possibilidade de ocorrência ou não de um risco, dependendo essencialmente do nível de vulnerabilidades dos ativos que enfrentam ameaças e, ainda, do nível de recursos das fontes de risco para explorá-los. Spiegel (1978, p. 08) aponta que "há sempre uma incerteza quanto à ocorrência ou não de um determinado evento", definindo probabilidade de um evento P(E) como a ocorrência de h maneiras diferentes desse evento, em um total de n maneiras possíveis, todas igualmente possíveis, portanto, a probabilidade $P(E) = h/n$".

Todos esses elementos confirmam que uma análise de riscos não é trivial, especialmente, no contexto de riscos relacionados aos sistemas de IA, tendo-se por premissa a Inteligência Artificial como não-coisa e o impacto de riscos tecnológicos, os quais podem advir do não entendimentos dos algoritmos (métodos e técnicas em sistemas de IA) e da tomada de decisão automatizada.

[8] Texto original: "occurrence or change of a particular set of circumstances".
[9] Texto original: "outcome of an event".
[10] Texto original: "elements, which, alone or in combination has the intrinsic potential to give rise to risk".

4 Os Riscos de não se Entender um Algoritmo e a Tomada de Decisão Automatizada

Em Freitas (2021) após explicar algoritmos trata-se do fato de que o desconhecimento sobre como os algoritmos funcionam pode levar a riscos, a saber: (i) julgar mal o "poder" do algoritmo, (ii) enfatizar demais a sua importância, (iii) pensar erroneamente que o algoritmo é um "agente" independente e isolado e, finalmente, (iv) não perceber como o "poder" pode ser realmente implementado por tecnologias e algoritmos. Por isso, é preciso compreender que os algoritmos operam sobre dados (bits) e podem realizar uma infinidade de tarefas, ações e tomada de decisão, de modo que o texto a seguir é uma atualização do texto originalmente publicado.

Problemas complexos não tem resposta binária (sim ou não) e por este motivo a tomada de decisão é sempre mais complexa do que "sim" ou "não", podendo incluir um valor de probabilidade entre 0 e 100, sendo esta probabilidade obtida por algoritmos em sistemas de IA. Inicialmente, há que se entender o que se pretende com a análise algorítmica sob o olhar da governança visando o entendimento ou revisão de tomadas de decisão automatizadas, inclusa nos instrumentos de regulação de sistemas de IA ou mesmo em legislações e regramentos de proteção de dados pessoais. Assim, pergunta-se: Espera-se que o algoritmo seja responsável? Ou espera-se que o algoritmo seja explicável? Dois caminhos com base em diferentes habilidades podem ser estabelecidos: a responsabilidade e a explicabilidade.

Como demonstrado por Daniel Neyland (2016), por meio de trabalho etnográfico, em um projeto de responsabilidade algorítmica, tornar algoritmos responsáveis muitas vezes significa literalmente mudá-los - tornando-os "responsáveis" no jargão etnomedológico. Tornar algo passível de prestação de contas (*accountability*) significa conferir qualidades que o tornam legível para grupos de pessoas em contextos específicos. Um algoritmo responsável é, portanto, literalmente diferente de um explicável ou inexplicável, visto que a transparência, por definição, altera as práticas que constituem um algoritmo (SEAVER, 2017, p. 6).

E esse é exatamente o ponto: as mudanças que a transparência necessita são mudanças que se desejar ter em um algoritmo. Este é um exemplo claro de como os diferentes esforços para representar um objeto são coordenados entre si e estão potencialmente em conflito. Transparência não é uma revelação do que sempre esteve lá, mas uma reconfiguração metódica da cena social que a altera em direção a fins específicos (SEAVER, 2017, p. 6).

Pode-se, portanto, relacionar tais habilidades: responsabilidade e explicabilidade, bem como transparência, com uma característica intrínseca dos algoritmos, qual seja a complexidade. Para tal, necessita-se ter conhecimento sobre a necessidade de abstração na fase de projeto de *software* (*design*).

O *design* (projeto) de um *software* depara-se na verdade com um objeto a ser entendido e decifrado: a complexidade; que por sua vez está relacionada com a simplicidade (OUSTERHOUT, 2018, p. 13) (BASS; CLEMENTS; KASMAN, 2003) e com o problema a resolver, tendo-se 02 (dois) caminhos possíveis (OUSTERHOUT, 2018, p. 13-14): a) eliminar a complexidade, tornando o algoritmo e, consequentemente, o código-fonte mais

simples e mais óbvio ou b) encapsular a complexidade, para que os programadores possam trabalhar em um sistema sem serem expostos a toda a sua complexidade de uma só vez. E, nesse ponto, o projeto de *software* encontra a abstração, visto que pensar por módulos é mais fácil do que pensar o todo (SCHOPENHAUER, 2005, p. 89) (OUSTERHOUT, 2018, p. 33).

Reforça-se que há que se cuidar para não incluir detalhes não importantes ou omitir detalhes realmente importantes, a exemplo dos vieses ou enviesamentos (*biases*). Caso existam dúvidas sobre o enviesamento de resultados obtidos a partir de determinado algoritmo, haverá obscuridade sobre muitas informações necessárias ao desenvolvimento do sistema, por exemplo: a base de dados, o processamento propriamente dito, as heurísticas e o modelo lógico matemático de funcionamento do algoritmo (FREITAS, 2021).

Por isso, muitos se perguntam como os algoritmos podem conter vieses (*biases*)? Os algoritmos são criados por seres humanos e podem assim ser igualmente tendenciosos a partir das bases de dados utilizadas nas etapas de treinamento e validação de modelos (matemáticos, estatísticos ou probabilísticos). Há que se lembrar que algoritmos andam de mãos dadas com a complexidade. Como explicado por Freitas e Barddal (2019, p. 111) "Na prática, o modelo realizará cálculos e fornecerá respostas objetivas, neutras e confiáveis às consultas realizadas". É para isso que um algoritmo e um programa de computador devem ser desenvolvidos e programados. Os autores relembram que "os computadores não têm preferências nem atitudes" (FREITAS; BARDDAL, 2019, p. 120), mas é necessário levar em consideração que sistemas de IA funcionam com a premissa de que o algoritmo "aprende" como se comportar baseado na experiência passada, que por sua vez são o *input* ao algoritmo e constituem as bases de dados. Mais importante, deve-se ter em mente que as decisões passadas são majoritariamente proferidas por seres humanos, que são potencialmente tendenciosos.

A seguir discute-se um entre os muitos aspectos da explicabilidade de algoritmos, em cenários de aplicação de sistemas de IA quando se necessita explicar a tomada de decisão algorítmica, ou seja, o caminho percorrido pelo algoritmo até um determinado resultado, sendo isso denominado de IA Explicável ou *AI Explainable* ou XAI.

4.1 *A Tomada de Decisão Automatizada e a Explicabilidade de Algoritmos em Sistemas de IA - IA Explicável ou AI Explainable ou XAI*

Freitas (2021) discutiu o outro lado da explicabilidade de algoritmos, quando tal entendimento é necessário frentes às responsabilidades (*accountability*), uma vez que algoritmos que constituem sistemas de IA apontam para decisões e decisões impactam vidas humanas. Nesse contexto, há juristas e cientistas que declaram os algoritmos sempre como caixas fechadas (*black-box*) a serem decifradas. Algoritmos *black-box* são aqueles nos quais não se tem acesso às informações do projeto interno (*design*) e da implementação, por exemplo, a linguagem de programação. O termo *black-box* se opõe aos algoritmos *white-box*, na qual toda a lógica e codificação estão abertos e podem ser estudados e analisados passo a passo. Há ainda os algoritmos denominados como *grey-box*, ou seja, quando se tem acesso a parte da lógica, do fluxo de dados, entre outros elementos. Agora realiza-se uma análise a partir da explicabilidade propriamente dita aplicada em algoritmos de sistemas de IA.

Mohd. Ehmer Khan e Farmeena Khan (2012, p. 12) classificam algoritmos pelo grau de explicabilidade a partir das técnicas aplicadas na análise, a saber:

- Técnica de teste de algoritmos *white-box*: É a análise ou investigação detalhada da lógica interna e da estrutura do código-fonte. Nessas técnicas, é necessário que o perito tenha total conhecimento do código-fonte;
- Técnica de teste de algoritmos *black-box*: É uma técnica de análise sem nenhum conhecimento do funcionamento interno da aplicação. Examina-se apenas os aspectos fundamentais do sistema, os quais podem ter pouca ou nenhuma relevância com a estrutura lógica interna do programa de computador;
- Técnica de teste de algoritmos *grey-box*: inicialmente deve-se explicar que os algoritmos *grey-box* são uma composição de algoritmos *white-box* e *black-box*. Para esses algoritmos, necessita-se aplicar técnicas para testar o aplicativo/programa de computador com conhecimento limitado do funcionamento interno e, por outro lado, há que se ter conhecimento dos aspectos fundamentais do programa de computador.

Entende-se, portanto, que a explicabilidade de um algoritmo não é trivial, podendo exigir conhecimentos técnicos desde a fase de projeto de um programa de computador (*software design*) até os requisitos de execução, bases de dados e resultados obtidos. Por isso, é muito importante avaliar os riscos do não entendimento de algoritmos, especialmente sobre a tomada de decisão automatizada.

E quando o tema envolve sistemas de IA, especialmente os que usam técnicas complexas, a exemplo de *Deep Learning*, há que se compreender que na Era das Não-Coisas existem algoritmos que são obscuros ou opacos, não intuitivos e de difícil compreensão aos seres humanos. Por isso, autores como Alves e Andrade (2022) vem discutindo a mudança de paradigma de *black-box* para a "caixa de vidro", conceito esse que engloba, de acordo com os autores "transparente, fácil de visualizar e entender – que contribui para a identificação de correlações indesejáveis, estabelecidas no interior do algoritmo, permitindo que desenvolvedores de um sistema rastreiem e corrijam falhas e vieses ali presentes.". E, ainda, "a "caixa de vidro" permite a verificabilidade, auditoria e apuração de responsabilidade quando a IA toma decisões ilegais" (ALVES; ANDRADE, 2022, p. 368). Os autores tratam sobre a *AI Explainable* ou IA Explicável ou XAI (*e**X**plainable **A**rtificial **I**ntelligence*) mostrando um caminho para auxiliar na redução da obscuridade de modelos algorítmicos, de modo que possam ser corrigidos ou mitigados os problemas de enviesamento (*bias*) (ALVES; ANDRADE, 2022, p. 352).

Em um primeiro momento, pode-se assumir que um sistema de IA tem por base algoritmos explicáveis, ou seja, os resultados da solução podem ser compreendidos por humanos contrapondo técnicas de Aprendizagem de Máquina do tipo "caixa preta", nas quais, nem mesmo os desenvolvedores conseguem explicar como ou por quais motivos o modelo alcançou um determinado resultado ou tomou uma decisão específica.

Neste sentido, reforça-se que a qualidade dos dados de treinamento, incluindo-se, portanto, a variedade, autenticidade e confiabilidade das fontes de coleta de dados e dos dados

propriamente ditos. Além disso, há que se considerar que os dados devem estar corretamente rotulados (quando necessário), devem ser localizáveis, acessíveis, reutilizáveis e imparciais (sem viés, o que depende de diversos fatores e esta discussão está além do escopo desse artigo).

Os estudos de Vilone e Longo (2020) e Saranya e Subhashini (2023) apresentam revisões sistemáticas de literatura (RSL) sobre o tema: *Explainable Artificial Intelligence*. Primeiramente, respeitando o sequenciamento temporal, Vilone e Longo (2020) buscaram definir algumas fronteiras no tema, considerando 350 artigos selecionados a partir da base do Google Acadêmico (*Google Scholar*) aplicando os argumentos de pesquisa: "*explainable artificial intelligence*", "*explainable machine learning*" e "*interpretable machine learning*". Os artigos foram então analisados a partir de suas próprias listas de bibliografias, visando recuperar outros estudos relevantes, totalizando 393 artigos. Os autores identificaram um acréscimo significativo de estudos a partir dos anos 2000, tendo o tema se tornado altamente relevantes após 2010. Isso devido ao rápido aumento na popularidade das técnicas de *Machine Learning* e, em particular, das técnicas de *Deep Learning*.

Classificaram os artigos em 04 (quatro) grupos distribuídos conforme quantidade de artigos e percentuais apontados, a saber (VILONE; LONGO, 2020, p. 4): a) *reviews on methods for explainability* (53 artigos - 13,5%); b) *notions related to the concept of explainability* (85 artigos - 21,6%); c) *development of new methods for explainability* (196 artigos - 49,9%) e d) *evaluation of methods for explainability* (59 artigos - 15,0%).

Vilone e Longo (2020) mencionam diversas áreas em que as aplicações de sistemas de IA se tornaram interessantes, se não necessárias: comércio eletrônico, jogos, saúde, visão computacional e aplicações no âmbito da justiça criminal. Explicam que a obscuridade de algoritmos criou a necessidade de arquiteturas XAI motivada principalmente por 03 (três) razões: (i) a demanda para produzir modelos mais transparentes; (ii) a necessidade de técnicas que permitam aos humanos interagir com os algoritmos; (iii) a exigência de confiabilidade das inferências algorítmicas (VILONE; LONGO, 2020, p. 2). Os autores mencionam que o artigo 22 do Regulamento Geral de Proteção de Dados da União Europeia (GDPR) define os direitos e obrigações diante da tomada de decisão automatizada, introduzindo o direito de explicação (*right of explanation*) ao dar aos titulares de dados (nomenclatura da LGPD) o direito de obter uma explicação sobre inferência(s) produzida(s) automaticamente por um modelo e, ainda, confrontar e desafiar uma recomendação associada à usa decisão automatizada, particularmente quando tal decisão pode afetar negativamente um indivíduo, seja legalmente, financeiramente, mentalmente ou fisicamente. Os autores entendem que o Parlamento Europeu tentou abordar o problema relacionado à propagação de inferências potencialmente tendenciosas para a sociedade, assumindo que um modelo computacional pode ter sido treinado a partir de dados tendenciosos e desequilibrados estatisticamente. Paralaelamente, pode-se mencionar o artigo 20 da LGPD, já comentado anteriormente.

Destaca-se que Vilone e Longo (2020, p. 5-7) ao analisarem os artigos do grupo *reviews* criam uma classificação hierárquica dos temas relacionadas à XAI e à interpretabilidade de técnicas de Aprendizagem de Máquina. Essa categorização demostra a complexidade da explicabilidade em si, relacionando parâmetros sobre: campos de aplicação, abordagem de

construção de métodos de explicabilidade projetados especificamente para explicar o processo inferencial de modelos, formatos de saída (*output*) de métodos de explicabilidade, por exemplo: formato gráfico (grafos, redes, árvores) ou esquemas baseados em regras, métodos projetados para explicar a lógica de dados e modelos baseados em conhecimento aplicados a um tipo específico de problema, nomeadamente regressão (*regression*) ou classificação (*classification*). Além disso, Arrieta et al. (2019, p. 11-12) listam e explicam diferentes técnicas para compor formatos de saída para explicação de um modelo a partir de algoritmos *black-box*, a exemplo de: *text explanations*, *visual explanations*, local explanations, *explanations by example*, *explanations by simplification* e *feature relevance explanations*. Apresentar cada uma das técnicas está além do escopo deste livro, mas recomenda-se a leitura do artigo Arrieta et al. (2019).

Especificamente sobre como tornar sistemas de IA explicáveis, Vilone e Longo (2020, p. 7-14, 72-81) fazem uma análise bem detalhada, a partir da RSL realizada, demostrando que existem 02 (dois) tipos de modelos dependendo do estágio em que se quer que a explicabilidade seja aplicada: a) métodos *ante-hoc* e b) métodos *post-hoc*, os quais por sua vez, atendem às seguintes categorias: b.1) *model-agnostic methods* e b.2) *model-specific methods*. Explicam os autores que os métodos *ante-hoc* geralmente visam considerar a explicabilidade de um modelo desde o início e durante o procedimento de treinamento do modelo para torná-lo naturalmente explicável, ao mesmo tempo que se busca atingir precisão ótima ou erro mínimo. Já métodos *post-hoc* visam manter um modelo treinado inalterado e explicar seu comportamento usando um explicador externo no momento do uso da base de dados de teste. Como *model-agnostic methods* entende-se os métodos que são aplicáveis a qualquer modelo baseado em *Machine Learning*. Esses métodos de XAI não consideram os componentes internos de um modelo, como pesos em redes neurais ou informações estruturais, portanto, podem ser aplicados a qualquer modelo do tipo *black-box*. Os modelos denominados *model-specific methods* são limitados a classes específicas de modelos, podendo ser aplicados, por exemplo, para interpretação de um modelo de regressão linear ou para interpretação de pesos em camadas (*layers*) de redes neurais. Todos os métodos dependem da construção de uma taxonomia[11] voltada à explicabilidade de sistemas de IA.

[11] Originalmente a taxonomia é definida como sendo a ciência de nomear, descrever e classificar organismos, incluindo todas as plantas, animais e microrganismos do mundo. Os taxonomistas usam de observações morfológicas, comportamentais, genéticas e bioquímicas para identificar, descrever e organizar espécies em classificações, considerando também aquelas que são novas para a ciência. Para maiores esclarecimentos, recomenda-se: https://www.britannica.com/science/taxonomy e https://www.sciencedirect.com/topics/computer-science/taxonomy-classification. Na área de Processamento de Linguagem Natural (*Natural Language Processing*), a taxonomia compreende o processo de classificação automática de conceitos por meio de uma estrutura hierárquica de *corpus* de texto. No que se refere à explicabilidade, a taxonomia pode ser combinada com métodos de Processamento de Linguagem Natural para tornar as informações utilizáveis tanto por pessoas quanto por computadores. Isso envolve analisar elementos significativos em um conteúdo (por exemplo em um texto, analisar verbos, substantivos, adjetivos, entre outros) por meio de uma análise estatística e identificar relacionamentos contextuais entre elementos distintos (por exemplo, palavras). A taxonomia permite fornecer um contexto hierárquico para conceitos e extrair as palavras usadas para descrevê-los. Simplificando, o termo taxonomia pode ser visto como sinônimo de tipologia, estrutura e, até mesmo, classificação, visto que taxonomias são formas úteis para representar conhecimento sobre objetos em um determinado domínio e para tal objetos de interesse precisam ser corretamente classificados. Para maiores esclarecimentos, recomenda-se: LANDOLT, Severin; WAMBSGANß, Thiemo; SÖLLNER, Matthias. A Taxonomy for Deep Learning in Natural Language Processing. In: Proc. Of 54th Hawaii International Conference on System Sciences (HICSS), 2021. p. 1061-1072. Disponível em: http://hdl.handle.net/10125/70741 Acesso em: 22 ago. 2024.

Arrieta et al. (2019, p. 19) apresentam uma taxonomia baseada em revisão de literatura para técnicas de explicabilidade relacionadas a diferentes modelos de *Machine Learning*. As técnicas XAI aplicadas na literatura referem-se ao uso de imagem, texto ou dados tabulados. Arrieta et al. (2019, p. 29) apresentam também uma taxonomia para modelos de *Deep Learning*, aprofundando a complexidade das discussões ao longo do trabalho. Analisando os diferentes modelos de *Machine Learning*, Arrieta et al. (2019, p. 29) discutem e demonstram que quanto mais interpretável é um modelo menos precisão (*Accuracy*) haverá na explicação do modelo. Portanto, interpretabilidade e precisão são inversamente proporcionais.

Mas para que explicar? Vilone e Longo (2020, p. 8-12) exploram 04 (quatro) motivos que suportam a necessidade de explicar a lógica de um sistema inferencial ou um algoritmo de aprendizagem, a partir da RSL realizada: a) justificar as decisões tomadas pela utilização de um modelo subjacente; b) controlar, aumentando a transparência dos modelos e seus respectivos funcionamentos, permitindo sua depuração e a identificação de falhas potenciais; c) melhorar a precisão e a eficiência de modelos; d) descobrir novos conhecimentos e estabelecer o aprendizado de relacionamentos e padrões.

Muitos são os métodos de explicabilidade que podem ser utilizados em sistemas de IA. Citar ou explicar tais métodos está além do escopo deste livro, mas recomenda-se a leitura para aprofundamentos (VILONE; LONGO, 2020). Os autores apontam que são vários os métodos para explicabilidade e que tais métodos podem funcionar com muitas técnicas de Aprendizagem de Máquina. No entanto, isso não significa que eles podem ser aplicados universalmente, pois podem ser limitados pelas entradas (*input*) da situação-problema que se busca resolver e pela explicação que se espera fornecer.

E para demonstrar a complexidade relacionada à explicabilidade de sistemas de IA, Vilone e Longo (2020, p. 9) apresentam todos os demais conceitos associados à explicabilidade, os quais foram mantidos em inglês para facilitar futuros estudos, estando tais conceitos listados em ordem alfabética como originalmente apresentados: *Algorithmic transparency, Actionability, Causality, Completeness, Comprehensibility, Cognitive relief, Correctability, Effectiveness, Efficiency, Explicability, Explicitness, Faithfulness, Intelligibility, Interactivity, Interestingness, Interpretability, Informativeness, Justifiability, Mental Fit, Monotonicity, Persuasiveness, Predictability, Refinement, Reversibility, Robustness, Satisfaction, Scrutability/diagnosis, Security, Selection/simplicity, Sensitivity, Simplification, Soundness, Stability, Transparency, Transferability, Understandability.* A explicação de cada conceito encontra-se tanto em Vilone e Longo (2020) quanto nos 393 artigos analisados pelos autores. Não se apresenta aqui este detalhamento, mas esta lista serve para deixar claro quanto a explicabilidade de sistemas de IA, tão falada e bradada ao vento por muitos que não são especialistas, é complexa e não trivial.

A segunda revisão sistemática de literatura (RSL) foi realizada por Saranya e Subhashini (2023) com objetivos a partir de 03 (três) focos de atenção: (i) artigos relacionados com XAI a partir de áreas de aplicação: agricultura, visão computacional (*Computer Vision*), finanças, previsão (*forecasting*), saúde, sensoriamento remoto e processamento de sinais, mídias sociais e transportes; (ii) conceitos, métodos, princípios e propriedades da explicabilidade; (iii) desafios em XAI. Foram analisados 91 artigos publicados entre janeiro de 2018 e outubro de 2022 e coletados a partir das seguintes bases de periódicos: Scopus, Web of Science, IEEE

Xplore e PubMed. Os 91 artigos foram selecionados de um total de 3545 artigos originalmente recuperados destas bases e a RSL guiou a metodologia para a seleção final de artigos.

Inicialmente, Saranya e Subhashini (2023, p. 4-80) mostram que a distribuição de artigos a partir da área de aplicação apresenta os seguintes resultados elencados em ordem decrescente: saúde – 43%, finanças – 9%, agricultura – 7%, visão computacional (*Computer Vision*) – 7%, mídias sociais – 7%, transportes – 7% e sensoriamento remoto e processamento de sinais – 2%. Ao longo do período de pesquisa o ano de 2022 representou 42% da amostra selecionada (38 artigos), 2021 – 31% (28 artigos), 2020 – 11% (10 artigos), 2019 – 11% (10 artigos) e 2018 – 5% (05 artigos), demonstrando assim o interesse pelo tema.

As associações entre temas e ideias relacionadas com explicabilidade existem e Saranya e Subhashini (2023, p. 7-8) apontam para 03 (três) grupos que podem ser formados a partir dos seguintes termos: Atributos por explicabilidade, Tipos de explicação e Estrutura de uma explicação. Estes grupos remetem a construção de uma taxonomia voltada à explicabilidade e dependente da área de aplicação de um sistema de IA. É fácil compreender que uma taxonomia construída para a área de finanças será diametralmente diferente de uma voltada à área de saúde, por exemplo. Isso devido a fatores como: estrutura da linguagem, palavras, termos e relações utilizadas na área específica de aplicação e, ainda, se a taxonomia terá um enfoque baseado em funções (*function-based approach*), resultados (*result-based approach*) ou conceitual (*conceptual approach*). Taxonomias baseadas em funções são construídas para identificar e classificar as funções básicas de um método de explicabilidade, extraindo informações sobre o modelo que integra o sistema de IA. Por outro lado, as taxonomias baseadas em resultados trabalham a partir dos resultado gerados por um modelo de explicabilidade, constituindo um importante componente do modelo de classificação. E, as taxonomias conceituais consideram como elementos-base: estágio em que se quer que a explicabilidade seja aplicada (*ante-hoc* e *post-hoc*), escopo (local ou global - local se refere à uma explicação para a predição a partir de uma determinada entrada e global se refere à uma explicação completa do modelo), tipo de problema (classificação ou regressão), dados de entrada – *input* (numérico, imagem, série temporal, texto, representação vetorial) e formato de saída dos resultados de explicação – *output* (numérico, regras, texto, visual, híbrido).

Saranya e Subhashini (2023, p. 11) e Phillips et al. (2021, p. 2-5) por meio do *National Institute of Standards and Technology* (NIST) enumeraram 04 (quatro) princípios que norteiam a Inteligência Artificial Explicável, a saber:

- Explicação (*Explanation*): os sistemas fornecem evidências ou razões que acompanham todas as saídas ou resultados (output) ou procedimentos;
- Significância (*Meaningful*): os sistemas fornecem explicações compreensíveis aos usuários de interesse;
- Precisão da explicação (*Explanation Accuracy*): a explicação reflete corretamente o procedimento aplicado pelo sistema de IA para geração das saídas ou resultados (output);

- Limites de conhecimento (*Knowledge Limits*): o sistema somente opera sob as condições para as quais foi projetado ou quando sua saída (output) tiver alcançado níveis suficientes de confiança.

Para Saranya e Subhashini (2023, p. 11) e Phillips et al. (2021, p. 2-5), o princípio da Explicação (*Explanation*) é o princípio base norteador dos demais princípios, visto que um sistema de IA precisa primeiramente ter uma explicação ou conter evidências que o acompanhem e que possam ser acessadas. Deve-se mencionar também que os princípios de Significância (*Meaningful*) e Precisão (*Accuracy*) referem-se à explicação propriamente dita, o que não verifica a qualidade da explicação. Deste modo, não é necessário explicar o processo executado pelo sistema relacionando-o diretamente com o resultado alcançado. Não se pode confundir precisão da explicação com precisão de decisão, visto que a avaliação da precisão de decisão envolve considerar se o sistema está acertando ou errando, portanto, decorre de conceitos estatísticos envolvendo comparações para medir o erro em torno do resultado (SPIEGEL, p. 300-307).

A RSL realizada por Saranya e Subhashini (2023, p. 11) mostra que a explicabilidade possui as seguintes propriedades, a saber: a) estilo: como a explicação é fornecida, podendo conter os seguintes elementos: a.1) nível de detalhamento da explicação; a.2) grau de interação humano-máquina; a.3) formato de saída dos resultados de explicação – *output*; b) propósito: motivos para que a explicação seja fornecida ao usuário.

Finalmente, Saranya e Subhashini (2023, p. 11-12) indicam os 03 (três) grandes desafios da XAI: Como criar modelos que sejam mais fáceis de explicar, como desenvolver interfaces de explicação e como compreender as condições psicológicas necessárias para explicações persuasivas. Percebe-se que a explicabilidade estabelecerá interconexão entre diferentes áreas do conhecimento científico, a exemplo de: linguagem natural, Psicologia, neurociências, Ciência da Computação, entre outras.

Phillips et al. (2021, p. 19) explicam que "*Humans are able to produce a variety of explanation types.*"[12]. E, alertam para o fato de que produzir explicações verbais é uma atividade cognitiva complexa para seres humanos e que tais explicações podem interferir nos processos de tomada de decisão e raciocínio humanos. E, é por isso que ao se ganhar experiência em um determinado tema, os processos subjacentes se tornam mais automáticos, fora da consciência e, portanto, mais difíceis de explicar verbalmente. Para o ser humano tudo isso passa em sua mente e inteligência como procedimentos que vão sendo refinados a cada exposição verbal. Por isso, um docente com 40 anos de experiência tem maior habilidade e automação nas explicações verbais. Mas todo esse processo gera tensões, basta ver um docente novato em sala de aula. Os autores afirmam que essa tensão está sendo repassada aos sistemas de IA, pois os seres humanos desejam alta precisão acompanhada do aumento da explicabilidade, o que pode ser de difícil alcance. Phillips et al. (2020, p. 19) invertem o jogo, afirmando que "*some assessments from humans may be more accurate when left automatic and implicit, compared to requiring an explicit judgment or explanation. Human judgments and decision making can oftentimes operate as a*

[12] Tradução livre: Os humanos são capazes de produzir uma variedade de tipos de explicação.

closed-box, and interfering with this closed-box process can be deleterious to the accuracy of a decision."[13]. Resumidamente, nem os seres humanos conseguem explicar por completo o caminho executado por sua própria tomada de decisão.

O estudo ora realziado trabalha com a premissa de que algoritmos e sistemas de IA são não-coisas, portanto, tornar transparente, interpretável e explicável, para os seres humanos, um conjunto de *bits* (0 e 1) é complexo, tanto do ponto de vista computacional quanto do ponto de vista do entendimento para seres humanos. Isto posto, uma vez que se espera que a transparência advenha da explicação desde os parâmetros dos modelos até a justificativa dos resultados. Arrieta et al. (2019, p. 10-11) pondera que existem 03 (três) níveis dentro da transparência a serem contemplados: transparência algorítmica, decomponibilidade e simulabilidade. Estes níveis são como círculos concêntricos, de modo que um modelo simulável é ao mesmo tempo um modelo que é decomponível e algoritmicamente transparente.

Inicialmente, Arrieta et al. (2019, p. 10-11) conceituam transparência algorítmica por meio da capacidade do usuário de entender o processo seguido pelo modelo para produzir qualquer saída (*output*) a partir de dados de entrada (*input*). Assim, ao se considerar um modelo linear, esse será transparente visto que seu espaço de erro (espaço representacional) pode ser entendido e raciocinado, permitindo que o usuário compreenda como o modelo agirá em todas as situações que o algoritmo enfrentará. Em aprendizagens profundas (*deep*) isso já não será possível, uma vez que o espaço de erro será opaco e não poderá ser totalmente observado, havendo a necessidade de a solução ser aproximada por meio de otimização heurística. A principal restrição para modelos algoritmicamente transparentes é que o modelo tem que ser totalmente explorável por meio de análise e métodos matemáticos.

Há que se ter muito cuidado ao utilizar o termo transparência algorítmica, visto que por vezes os autores incorrem em erro, generalizando todos os conceitos ora apresentados e discutidos como se assim o fosse. Nada é trivial, está a se falar de complexidade e, portanto, cabe o cuidado dos autores que não dominam a área tecnológica.

Para Arrieta et al. (2019, p. 10-11) o segundo nível dentro da transparência é a decomponibilidade a qual se refere a capacidade de explicar cada uma das partes de um modelo (entradas - *inputs*, parâmetros e cálculos). Por isso é decomponível, em partes, podendo ser sinônimo de inteligibilidade. Este nível relaciona-se com a capacidade de entender, interpretar ou explicar o comportamento de um modelo. No entanto, como ocorre com a transparência algorítmica, nem todo modelo pode cumprir essa propriedade. Decomponibilidade requer que cada parte do modelo seja interpretável. A restrição adicional para que um modelo algoritmicamente transparente se torne decomponível é que cada parte do modelo deve ser compreensível por um humano sem a necessidade de ferramentas adicionais, o que pode dificultar ou inviabilizar a transparência.

[13] Tradução livre: algumas avaliações de humanos podem ser mais precisas quando deixadas automáticas e implícitas, em comparação à exigência de um julgamento ou explicação explícita. Os julgamentos e a tomada de decisões humanas podem muitas vezes operar como uma caixa fechada, e interferir nesse processo de caixa fechada pode ser prejudicial à precisão de uma decisão.

E, o nível mais externo dentro da transparência é simulabilidade, de acordo com Arrieta et al. (2019, p. 10-11). A simulatabilidade denota a capacidade de um modelo de ser simulado ou pensado por um humano, portanto, a complexidade assume um lugar dominante neste nível. Sistemas baseados em regras simples, mas com grande quantidade de regras, não se beneficiam deste nível de transparência. Já uma Rede Neural poderá contar com a simulabilidade, visto que um modelo interpretável é aquele que pode ser facilmente apresentado a um humano por meio de texto e visualizações, para que o humano pense e raciocine sobre o modelo como um todo.

A interpretabilidade precisa ir além, para que o modelo possa ser compreendido pelos seres humanos, ou seja, há necessidade de entendimento sobre como o modelo toma decisões. Já a explicabilidade refere-se a tudo que foi apresentado, tendo-se como ressalva à capacidade de decifrar por que um resultado foi calculado e obteve-se um determinado valor considernado-se a base de dados ou uma situação específica apresentada ao sistema de IA.

Muitas são as representações gráficas ou textuais que buscam "desenhar" ou "explicar" sistemas de IA. Pode parecer paradoxal, mas explicar, na prática, significa também ter uma representação de como a XAI pode seguir um conjunto de passos ou estabelecer um método genérico, em alto nível de representação aos seres humanos. Há representação baseada no fluxo de dados, na interface com o usuário (*User Interface* – UI), em listas de perguntas e/ou respostas, entre outras (ARRIETA et al., 2019, p. 4-10; GUNNING et al., 2019, p. 5; WACHTER et al., 2018, p. 843). Ao que se refere à compreensão humana, o foco está em como as pessoas raciocinam, considerando seus erros e, ainda, como as pessoas deveriam raciocinar e explicar sua tomada de decisão. Já do ponto de vista da IA Explicável, há que se compreender como a XAI suporta raciocínios, considerando erros e, ainda, buscar respostas para como a XAI pode gerar explicações (WANG et al., 2019, p. 4).

Finalmente, pode-se apontar que um sistema de IA responsável necessita de projeto, desenvolvimento e implementação a partir de valores éticos, incluindo-se a confiabilidade, transparência, explicabilidade, justiça, robustez, proteção de dados pessoais, não violação aos direitos fundamentais e respeito aos Direitos Humanos. A tão desejada XAI é elemento primordial para que a IA possa conquistar a confiança e a segurança necessárias no mercado e entre seres humanos para, então, fomentar sua adoção baseada em benefícios. Conquistar confiança é componente primordial de uma relação humana e, também, será em relações entre humanos e máquinas. Confiança integra a principiologia jurídica e pertence ao núcleo da ordem jurídica, portanto, opera em conjunto com outros princípios a exemplo da boa-fé e da segurança jurídica. Enfim, a temática não é trivial.

5 Conclusão

O artigo trouxe definições de sistemas de Inteligência Artificial e partiu da premissa que a IA depende de 02 (dois) elementos que são a base do seu funcionamento: (i) algoritmos e (ii) dados; portanto, bits (*binary digit*), conjuntos de zeros (0) e uns (1). Isso levou o estudo a assumir a IA como não-coisa com base na filosofia de Byung-Chul Han (2022) sobre não-coisas e, assim, fomentar uma discussão em torno de riscos decorrentes de sistemas de IA,

trabalhando especialmente os riscos de não se entender um algoritmo e a tomada de decisão automatizada. Apresentou-se e trabalhou-se a necessidade de se alcançar a explicabilidade em sistemas de IA, ou IA Explicável ou *AI Explainable* ou XAI, passando por responsabilidade, complexidade, verificabilidade e transparência.

Escalar o uso de sistemas de IA permitirá ir além de operações manuais e automatização tradicional. Existem tarefas e procedimentos complexos que podem se beneficiar da aplicação de sistemas de IA. A confiança virá atrelada à segurança, visto que modelos seguros tanto do ponto de vista tecnológico quanto jurídico aumentarão a velocidade de adoção da IA e os benefícios ficarão cada vez mais evidentes. E a confiança e a segurança virão com a explicabilidade, uma vez que entende-se ser imperativo explicar como um sistema de IA alcança uma decisão, como um resultado é gerado e como esse resultado é afetado pelo modelo treinado por uma base de dados confiável e, ainda, os motivos (caminhos internos) que ativaram o sistema de IA e o fez atuar em determinada situação real.

A explicabilidade poderá dar à luz (nascimento) aos sistemas de IA em grande escala, destacando dados, modelos e procedimentos por meio de análises transparentes como vidro e com validade não somente matemática, estatística ou probabilística, mas para seres humanos. Espera-se que a explicabilidade possa evidenciar falhas, vulnerabilidades e vieses, mitigando riscos e permitindo a adequação (ou até mesmo a correção) dos sistemas de IA.

A transparência será a chave de toda a explicabilidade. Urge pensar em interface de usuários para que os humanos possam "observar" os diferentes níveis, como se subissem uma escada, um degrau de cada vez. Pular do primeiro degrau ao último pode ser desastroso. Treinar modelos para sistemas de IA não é uma tarefa trivial, como já comentado anteriormente, mas é preciso estabelecer um ciclo entre céticos e aqueles que acreditam na IA, entre experientes e novatos, para que os algoritmos sejam aprimorados em termos de aprendizado (humano e algorítmico), produtividade, confiança e transparência.

Lembre-se que algumas perguntas podem guiar a decisão humana pelo uso de sistemas de IA, bem como averiguar se tem-se uma XAI realmente aplicável em termos de devida diligência, quando se pensa em responsabilidade, transparência e prestação de contas e visa-se mitigar riscos, a exemplo de: Qual a formação da base de dados? A base foi pré-processada, houve identificação de enviesamentos? Como a base de dados foi formada: coleta de dados, dados já existentes em outros sistemas, inclui dados pessoais, dados de geolocalização, cadastro em sites ou aplicativos, etc)?; Quais técnicas/algoritmos de IA com compõem o sistema?; Como o(s) modelo(s) são treinados, validados e testados?; Há um processo contínuo de treinamento?; Como está estruturada a explicabilidade do sistema?; Há explicações para os resultados?; Há recomendações a partir dos resultados?; Quais ações podem ser tomadas a partir dos resultados?; Já foram detectados vieses?; Tais vieses foram eliminados?; Como foram eliminados?; Houve eliminação ou somente foram minorados os efeitos do enviesamento?; O sistema de IA conta com um processo contínuo de aprimoramento?; Há interferência humana ou validação por humanos?; O sistema de IA foi desenvolvido para avançar progressivamente de maneira automática?. Estas e muitas outras perguntas podem ser formuladas, bem como muito pode ser aprofundado sobre o tema de XAI.

Por tudo isso a obscuridade de sistemas de IA não só torna as decisões mais difíceis de serem entendidas, ou corretamente explicadas, mas pode causar impactos sobre os indivíduos afetados pelos sistema, visto que ocultar falhas em sistemas de IA, como a existência de imprecisões ou vieses, gera riscos de difícil mitigação.

E, por último, e não menos importante, cabe apontar que não haverá humanos revisando a tomada de decisão, mas algoritmos revisando algoritmos. Diante da IA como não-coisa, cai por terra a premissa de que as revisões de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais, artigo 20 da LGPD (BRASIL, 2018), devem ser realizadas por humanos. Se os sistemas regulatórios de IA apostarem em estruturas XAI poder-se-á revisar não somente as decisões tomadas unicamente com base em tratamento automatizado de dados pessoais, mas todas as decisões automatizadas que afetem os interesses dos titulares de dados, usuários-consumidores-cidadãos, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade. Existirá a caixa transparente proposta por Arrieta et al. (2019) ou a caixa de vidro imaginada por Alves e Andrade (2022).

6 Referências

ABNT ISO/IEC Guia 73. **Gestão de riscos – vocabulário: recomendações para uso em normas**. 2005.

ABNT. Associação Brasileira de Normas Técnicas. **ABNT NBR ISO/IEC - 22989:2023** – Tecnologia da Informação – Inteligência Artificial – Conceitos de Inteligência Artificial e Terminologia. 2023.

AGÊNCIA DOS DIREITOS FUNDAMENTAIS DA UNIÃO EUROPEIA. **Preparar o futuro – Inteligência artificial e direitos fundamentais: síntese**, Luxemburgo: Serviço das Publicações da União Europeia, 2021.

ALVES, M. A.; ANDRADE, O. Da "caixa-preta" à "caixa de vidro": o uso da explainable artificial intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Direito Público**, v. 18, n. 100, out-dez, 2021. p. 349-373.

ARRIETA, A. et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. **Information Fusion**, v. 58, p. 82.115, 2019. Disponível em: https://api.semanticscholar.org/CorpusID:204824113 Acesso em: 17 dez. 2024.

BASS, Len; CLEMENTS, Paul; KAZMAN, Rick. **Software architecture in practice**, 2nd edition, Addison-Wesley, 2003.

BECK, U. Conversation 3: global risk society. In: BECK, Ulrick; WILLMS, Johannes (Org.). **Conversations with Ulrich Beck**. Trad. de Michael Pollak. Cambridge: Polity, 2004.

BECK, U. **La sociedad del riesgo: hacia una nueva modernidad**. Trad. de Jorge Navarro, Daniel Jiménez, Maria Rosa Borrás. Barcelona: Paidós, 1998.

BERNSTEIN, P. L. **Against the gods: the remarkable story of risk**. New York: John Wiley & Sons Inc., 1996.

BRASIL. **Projeto de lei n° 2338**, de 2024. Dispõe sobre o uso da Inteligência Artificial. Complementação de voto. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=9683716&ts=1720704050239&rendition_principal=S&disposition=inline Acesso em: 17 dez. 2024.

BRASIL. **Lei 13.709**, de 14 de agosto de 2018, Lei Geral de Proteção de Dados - LGPD, 2018.

BRIN, D. **The transparency society**: will technology force us to choose between privacy and liberty? United States: Perseus Book, 1998.

CALDAS, A; FREIRE, V. **Cibersegurança: das preocupações à ação**. Instituto de Defesa Nacional – IDN, Working Paper 2, Lisboa, Portugal, 2013.

CAVEDON, R; FERREIRA, H. S.; FREITAS, C. O. A. O Meio Ambiente Digital sob a Ótica da Teoria da Sociedade de Risco: Os avanços da informática em debate. **Revista Direito Ambiental e Sociedade**, v. 5, p. 194-223, 2015.

CNIL. **Privacy impact assessment (PIA): methodology**. French Data Protection Authority, 2018.

DATA PROTECTION WORKING PARTY. **Guidelines on data protection impact assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679 (WP29)**. Artic. 29 Data Prot. Work. Party. WP 248 rev 22 (2017). 2017.

FLORIDI, L. **The 4th Revolution**: How the Infosphere is Reshaping Human Reality. New York: Oxford University Press, 2014.

FREITAS, C. O. A. O direito e a inteligência artificial como não-coisa. **Conpedi Law Review**, Florianopolis, Brasil, v. 10, n. 1, 2024. DOI: 10.26668/2448-3931_conpedilawreview/2024.v10i1.10710.

FREITAS, C. O. A. Riscos e proteção de dados pessoais. **RRDDIS Revista Rede de Direito Digital, Intelectual & Sociedade**, v. 2, p. 225-247, 2023.

FREITAS, C. O. A. **A obscuridade dos algoritmos e a revisão da tomada de decisão automatizada diante de segredos comerciais e industriais**. In: Marcos Wachowicz; Marcelle Cortiano. (Org.). Sociedade informacional & propriedade intelectual. 1ed.Curitiba: GEDAI Publicações/UFPR, 2021, v. 1, p. 221-245.

FREITAS, C. O. A.; BARDDAL, J, P. Análise preditiva e decisões judiciais: controvérsia ou realidade?. **Democracia Digital e Governo Eletrônico**, 2019, v. 1, p. 107-126.

GELLERT, R. Understanding the notion of risk in the General Data Protection Regulation. **Computer Law & Security Review: The International Journal of Technology Law and Practice**, p. 1-10, 2017.

GODARD, O. et al. **Traité des nouveaux risques**, Paris: Éditions Gallimard, 2002.

GUNNING, D. et al. DARPA's explainable AI (XAI) program: A retrospective. **Applied AI Letters**, v. 2, issue 4, Special issue: DARPA's Explanable Artificial Intelligence (XAI) Program, December, 2021. p. 1-11. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1002/ail2.61 Acesso em: 17 dez. 2024.

HAN, B-C. **Não-coisas**: transformações no mundo em que vivemos. Trad. Ana Falcão Bastos. Lisboa: Relógio D'Água Editores, 2022.

HARARI, Y. N. **Homo Deus**: uma breve história do amanhã. Trad. Paulo Geiger. 1ª. ed., São Paulo: Companhia das Letras, 2016.

HLEG - High-Level Expert Group on Artificial Intelligence. **A definition of AI**: Main capabilities and scientific disciplines. European Commission, Brussels, 2019.

ISO/IEC 27001. I**nformation technology — security techniques — information security management systems — requirements**. 2022.

ISO. **Risk management - principles and guidelines**. International Organization for Standardization, Geneva, Switzerland, 2009.

KHAN, M. E; KHAN, F. A comparative study of white box, black box and grey box testing techniques. international journal of advanced, **Computer Science and Applications - IJACSA**, Vol. 3, No. 6, 2012. p. 12-15.

NEYLAND, D. Bearing account-able witness to the ethical algorithmic system**. Science, Technology & Human Values**, vol. 41, issue 1, 2016, p.50–76. Disponível em: https://www.researchgate.net/publication/282803178_Bearing_Accountable_Witness_to_the_Ethical_Algorithmic_System Acesso em: 17 dez. 2024.

OUSTERHOUT, John. **A philosophy of software design**. Yaknyam Press: Palo Alto, CA, USA, 2018.

OXFORD LIVING DICTIONARIES, Artificial intelligence, 2023. Disponível em: https://www.oed.com/viewdictionaryentry/Entry/271625 Acesso em: 17 dez. 2024.

PARDO, José Esteve. **Técnica, riesgo y derecho**. Barcelona: Ariel, 1999.

PHILLIPS, J. et al. NISTIR 8312 - Four Principles of Explainable Artificial Intelligence. U.S. Department of Commerce, **NIST - National Institute of Standards and Technology**, 2021.

SARANYA, A.; SUBHASHINI, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. **Decision Analytics Journal**, v. 7, january, 2023. p. 1-14.

SCHOPENHAUER, Arthur. **O mundo como vontade e como representação**. Trad. Jair Barbosa, São Paulo: UNESP, 2005.

SCHUILENBURG, M.; PEETERS, R. **The Algorithmic Society Technology**, Power, and Knowledge. United Kingdom: Routledge, 2021.

SEAVER, N. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. **Big Data & Society**, July–December, 2017, p. 1–12. Disponível em: https://journals.sagepub.com/doi/full/10.1177/2053951717738104 Acesso em: 17 dez. 2024.

SHENK, D. **Data Smog**: Surviving the Information Glut. Abacus, 1997.

SPIEGEL, M. R. **Probabilidade e estatística**. Coleção Schaum. Trad. Alfredo Alves de Farias. São Paulo: McGraw-Hill do Brasil, 1978.

UNIÃO EUROPEIA. **Artificial Intelligence Act**. Consolidated version. Bruxelas, 21.4.2021 COM (2021) 206, final P9_TA (2023)0236. Disponível em: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html Acesso em: 17 dez. 2024.

UNIÃO EUROPEIA. General Data Protection Regulation, 2016. Disponível em: https://gdpr-info.eu/ Acesso em: 17 dez. 2024.

VILONE, G.; LONGO, L. **Explainable artificial intelligence: a systematic review**. 2020. p. 1-81. Disponível em: https://arxiv.org/abs/2006.00093 Acesso em: 17 dez. 2024.

WACHTER, S. et al. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, **Harvard Journal of Law & Technology**, v. 31, n. 2,2018. p. 841-887. Disponível em: https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf Acesso em: 17 dez. 2024.

WANG, D. et al. Designing theory-driven user centric explainable AI. In: **Proceedings of the 2019 CHI - Conference on Human Factors in Computing Systems**, 601, 02 May 2019. p. 1–15. Disponível em: 10.1145/3290605.3300831

ZUBOFF, S. **A Era do Capitalismo de Vigilância**. Trad. George Schlesinger. Rio de Janeiro: Editora Intrínseca Ltda., 2021.