UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PAULO HENRIQUE PADOVAN

MAIA: Metamodelo de *Accountability* para Inteligência Artificial

Recife
2023

PAULO HENRIQUE PADOVAN

MAIA: Metamodelo de *accountability* para Inteligência Artificial

Tese/Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de doutor em Ciência da Computação. Área de concentração: Inteligência Computacional.

Orientador: Ruy José Guerra Barretto de Queiroz

Coorientadora: Clarice Marinho Martins

Recife

2023

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

P124m Padovan, Paulo Henrique
MAIA: Metamodelo de *Accountability* para Inteligência Artificial / Paulo Henrique Padovan – 2023.
333 f.: il., fig., tab.

Orientador: Ruy José Guerra Barretto de Queiroz.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2023.
Inclui referências e anexo.

1. Inteligência computacional. 2. Responsabilidade algorítmica. 3. IA confiável. 4. Inteligência Artificial Explicável (IAE). 5. Unidade Responsiva Explicável (URE). I. Queiroz, Ruy José Guerra Barretto de (orientador). II. Título

006.31 CDD (23. ed.) UFPE - CCEN 2023 – 108

Paulo Henrique Padovan

“**MAIA - Metamodelo de accountability para Inteligência Artificial”**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 08/02/2023

_____

Orientador: Prof. Dr. Ruy José Guerra Barretto de Queiroz

BANCA EXAMINADORA

_____

Prof. Dr. Frederico Luiz Gonçalves de Freitas
Centro de Informática / UFPE

_____

Profa. Dra. Jessyka Flavyanne Ferreira Vilela
Centro de Informática / UFPE

_____

Profa. Dra. Maria Amália Oliveira de Arruda Camara
Faculdade de Ciências da Administração e Direito de Pernambuco / UPE

_____

Prof. Dr. Sergio Torres Teixeira
Departamento de Direito Público Geral e Processual / UNICAP

_____

Prof. Dr. Virgilio Augusto Fernandes Almeida
Departamento de Ciência da Computação / UFMG

A meus pais pelos caminhos que me indicaram e incentivaram.

A minha família pela compreensão nas horas de ausência.

AGRADECIMENTOS

A Deus, aos meus orientadores Ruy José Guerra Barreto de Queiroz e Clarice Marinho Martins, aos meus pais, irmãos, sobrinha e a toda minha família que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa de minha vida, aos membros da banca, em especial ao presidente Frederico Luiz Gonçalves de Freitas, cujas valorosas contribuições enaltecem esta obra, e a todos que direta ou indiretamente colaboraram na elaboração desse trabalho.

“They all think it’s about more detail. But that’s not how memory works. We recall with our feelings. Anything real should be a mess.” (Blade Runner, 1982).

RESUMO

A inteligência artificial (IA) pode tomar decisões erráticas, causando preconceito, desconfiança, perdas ou danos aos indivíduos. Questões complexas de caráter moral e normativo devem ser dirimidas para estabelecer como determinar accountability. Até recentemente, entender o funcionamento das “caixas pretas” era extremamente difícil; no entanto, o uso de Inteligência Artificial Explicável (XAI) ajuda a aclarar os problemas complexos que podem incorrer a IA. Neste contexto, esta tese procura analisar, caracterizar e conformar accountability no âmbito da moral e da IA e fornecer um cânone que ajude as várias partes interessadas a lidar com as questões de responsabilidade algorítmica. Para definir accountability, apresentamos, caracterizamos e diferenciamos os diversos matizes de responsabilidade no contexto normativo atual, apresentando os dois pressupostos relacionados à accountability moral: exigibilidade e responsividade. Por fim, fornecemos uma análise sobre como a estrutura contábil existente, com o suporte de XAI e dados registrados, pode abordar questões relacionadas à responsabilidade algorítmica da IA.

Palavras-chave: inteligência artificial; responsabilidade algorítmica; IA confiável; Inteligência Artificial Explicável (IAE); Unidade Responsiva Explicável (URE).

ABSTRACT

Artificial intelligence (AI) can lead to erratic behavior causing bias, mistrust, loss, or damage to individuals. Intricate questions of moral and normative nature must be resolved to establish how to determine accountability. Until recently, understanding the inner workings of “black boxes” has been exceedingly difficult; however, the use of Explainable Artificial Intelligence (XAI) would help simplify the complex problems that can occur with AI. In this context, this thesis seeks to analyze, characterize, and conform accountability within the scope of morality and AI and provide a canon that helps the various stakeholders to deal with the urges for algorithmic responsibility. To define accountability, we present, characterize, and differentiate concepts of responsibility, liability, and accountability in today's normative context, presenting the two assumptions related to moral accountability: enforceability and answerability. Lastly, it provides an analysis of whether existing accounting framework, with the support of XAI and logged data, can address concerns related to AI algorithmic accountability.

Keywords: artificial intelligence; algorithmic accountability; trustworthy AI; Explainable Artificial Intelligence (XAI); Explainable Accountable Unit (XAU).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAA19	*Algorithmic Accountability Act of* 2019
AAA22	*Algorithmic Accountability Act of* 2022
ACCV	Análise de Custo do Ciclo de Vida
ACR	Análise da Causa Raiz
ACV	Análise do Ciclo de Vida
AFD	Analista Forense Digital
AIGO	*Expert Group on AI in Society*
ANPD	Autoridade Nacional de Proteção de Dados
ATEAC	*Advanced Technology External Advisory Council*
BIA	Bradesco Inteligência Artificial
BPMN	*Business Process Model and Notation*
Brexit	*British Exit*
CA	*Cambridge Analytica*
CFR	*Code of Federal Regulations*
CNN	*Convolutional Neural Network*
COMPAS	*Correctional Offender Management Profiling for Alternative Sanctions*
CPDL	*Controller Pilot Data Link*
CRISP-DM	*CRoss-Industry Standard Process Model*
CRM	*Customer Relationship Management*
CVR	*Cockpit Voice Recorder*
DEG	Distância Entre Grupos
DETRAN	Departamento Estadual de Trânsito
DGW	Distância de Gromov-Wasserstein
DICS	Dados–Informação–Conhecimento–Sabedoria
DIG	Distancia Intragrupo
DIKW	Data–Information–Knowledge–Wisdom
DSRPAI	*Dartmouth Summer Research Project on Artificial Intelligence*
DW	Distância de Wasserstein
EAD	Ensino a Distância
EBIA	Estratégia Brasileira de Inteligência Artificial
EC	Espaço Construto
ECU	*Electronic Control Unit*
ED	Espaço Decisório
EDR	*Event Data Recorder*
ENEM	Exame Nacional do Ensino Médio
ENIAC	*Electronic Numerical Integrator and Calculator*
EO	Espaço Observado
ETC	Extrair Transformar e Carregar
Eurocae	*European Organization for Civil Aviation Equipment*
FAA	*Federal Aviation Administration*
FD	Forense Digital
FDR	*Flight Data Recorders*
FMEA	*Failure Mode and Effects Analysis*
FMECA	*Failure Mode Effects and Criticality Analysis*
FRAM	*Functional Resonance Analysis Method*
FTA	*Fault-Tree Analysis*
FTC	*Federal Trade Commission*

GCS	Gerenciamento da Cadeia de Suprimentos
GDPR	*General Data Protection Regulation*
GESI	Gerenciamento de Eventos de Segurança da Informação
GOFAI	*Good Old-Fashioned Artificial Intelligence*
GRC	Gerenciamento de Relacionamento com o Cliente
IA	Inteligência Artificial
IAE	Inteligência Artificial Explicável
IAEF	Inteligência Artificial Explicativa Forense
ICAO	*International Civil Aviation Organization*
ICE	*Individual Conditional Expectation*
IDEF0	*ICAM DEFinition for Function Modeling*
IEEE	*Institute of Electrical and Electronic Engineer*
IPIA	Influência, Previsibilidade, Interesse e Atitude
IRAR	Identificação; Registro; Análise e Relato das informações
ISO	*International Organization for Standardization*
KDD	*Knowledge Discovery in Databases*
KPI	*Key Performance Indicator*
LAW	*Lethal Autonomous Weapon*
LGPD	Lei Geral de Proteção de Dados
LIDAR	*Light Detection And Ranging*
LIME	*Local Interpretable Model-Agnostic Explanations*
LSI-R	*Level of Service Inventory Revised*
MA	Modelos de Atividade
MAIA	Metamodelo de *accountability* para Inteligência Artificial
MCU	*Media Control Unit*
MD	Modelos de Dados
MDS	*Multidimensional Scaling*
ME	Modelos Estruturais
MIJ	Mecanismo Individualmente Justo
MIT	*Massachusetts Institute of Technology*
ML	*Machine Learning*
MN	Modelos de Negócios
MPN	Modelos de Processos de Negócios
MSJ	Mecanismo Socialmente Justo
NCA	*New Criminal Activity*
NTSB	*National Transportation Safety Board Office of Highway Safety*
NVCA	*New Violent Criminal Activity*
OCDE	*Organisation de Coopération et de Développement Économiques*
PC	Plano de Comunicação
PII	*Personal Identifiable Information*
PL	Projeto de Lei
PSA	*Public Safety Assessment*
QN1	Por que coletar e divulgar informações?
QN2	Quem são os stakeholders contemplados por essas informações?
QN3	Quais tipos de informações serão coletadas e quais divulgações serão feitas?
QN4	Como as informações devem ser divulgadas?
RCM	*Restraint Control Module*
RGPD	Regulamento Geral sobre a Proteção de Dados
RL	*Robot Laws*

SAS	*Statistical Analysis System*
SDA	Sistema de Decisão Automatizada
SDAAR	Sistema de Decisão Automatizada de Alto Risco
SE	Sistema Empresarial
SI	Sistema de Informação
SIC	Sistema de Informação Contábil
SPRE	Sistema de Planejamento de Recursos Empresariais
STAMP	*System Theoretic Accident Model and Processes*
TAY	*Thinking About You*
UE	União Europeia
UML	*Unified Modeling Language*
VG	Viés de Grupo
VPN	*Virtual Private Network*
XAI	*Explainable Artificial Intelligence*
XAU	*Explainable Accountable Unit*
XNN	*Explainable Neural Network*

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A cada dia nos maravilhamos com robôs, e.g., Alexa, Siri, GPT-3, capazes de nos compreender e executar pequenas tarefas para nós. Claro, nem tudo é perfeito. A Siri nem sempre tem resposta para nossas perguntas e a Alexa ocasionalmente nos espiona[1]. Também nos deparamos com robôs que em menos de 24 horas se tornam racistas e começaram a postar uma série de comentários ofensivos no Twitter[2].

Alguns desses robôs são responsáveis por filtrar as informações disponíveis, em pequenos sumários particularizados para nossas preferências. Mas, e quando esses robôs passam a difundir notícias falsas e a influir diretamente no resultado de pleitos como o *Brexit* e a eleição americana em 2016[3]?

Por falar em interferir, estes autômatos já se imiscuíram no judiciário. Na contemporaneidade esses mecanismos promovem redução de custos; aumento de produtividade e eficiência; democratização dos serviços e transparência; desafogam o judiciário; e diminuem o litígio por meio de mecanismos que auxiliam o judiciário em tarefas como: *analytics* e jurimetria; automação e gestão de documentos; conteúdo jurídico; educação e consultoria; extração e monitoramento de dados públicos; gestão (e.g., escritórios e departamentos jurídicos); AI (i.e., soluções de inteligência artificial para tribunais e poder público); redes de Profissionais; resolução de conflitos online; e métodos baseados em evidências (i.e., avaliadores de risco) (CÂMARA, 2018). Contudo, quando esses sistemas possuem vícios, os resultados podem ser catastróficos, variando de simples atrasos até indeferimento de condicional, e.g., *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS)[4].

---

[1] Vide (ARBULU, 2019) e (AMENO e DIAS, 2022).
[2] Vide Microsoft Tay item 4.3 Reputação.
[3] Vide *Bexit* item 4.6 Políticas
[4] Vide item 4.5.1.2 COMPAS

Constatamos que, à medida que a inteligência artificial evolui, cresceram as controvérsias causadas pelo seu uso. É curioso notar, no entanto, que toda essa celeuma, não é recente. Principalmente no tocante a automação.

Os primórdios da automação datam da revolução industrial e o advento das máquinas industriais entre 1790 e 1840[5]. Outrora, assim como hoje, as pessoas temeram por seus empregos ante a crescente onda de automação que varreu a Inglaterra, a Europa e o Mundo. Embora tudo tenha acabado bem para os trabalhadores – a exceção de uma revolução comunista e duas grandes guerras – essa primeira leva de automação ainda não nos proveu a IA, mas o que faltava para sua gênese, o computador.

Em 1837, Charles Babbage começou a desenvolver o protótipo de uma máquina, The Analytical Engine, que foi o primeiro dispositivo a ganhar o epíteto de computador. Concomitantemente, em 1842, Ada Lovelace criou os primeiros programas de computador que poderiam rodar na máquina. Infelizmente, Babbage morreu antes que seu protótipo estivesse completo[6].

Coube a pioneiros como: Alan Turing – apresentar o conceito de "máquina universal", capaz de computar qualquer coisa que seja computável, em 1936; John Atanasoff – projetar o primeiro dispositivo capaz de armazenar informações em sua memória principal, em 1937; John Mauchly e John Eckert – construirem o Electronic Numerical Integrator and Calculator (ENIAC[7]), em 1946; William Shockley, John Bardeen e Walter Brattain – inventarem o transistor, em 1948; Jack Kilby e Robert Noyce – revelarem o circuito integrado, em 1968, i.e., chip de computador[8]; até chegarmos aos novos desbravadores que desenvolvem a computação quântica, em 1985[9], e a molecular, i.e., moléculas como computadores, em 1994[10].

Tautócrono ao desenvolvimento da computação, mas sem a mesma graça, os conceitos de robô, automação e IA deram seus primeiros passos. O estudo da IA iniciou-se nos anos 50 com o *Logic Theorist* dos cientistas Allen Newell, Cliff Shaw e

---

[5] Vide (SULLIVAN, 2017).
[6] Vide (CERUZZI, 2003).
[7] A máquina contava com um hardware equipado com 70 mil resistores e 18 mil válvulas a vácuo.
[8] ibid.
[9] Vide (RIEFFEL e POLAK, 2014).
[10] Vide (VASILAKOS e CHANG, 2014).

Herbert Simon. O programa foi projetado para imitar as habilidades de resolução de problemas de um ser humano. Coube, no entanto, a John McCarthy cunhar o termo *Artificial Intelligence* (AI), Inteligência Artificial em português, em 1956 na *Dartmouth Summer Research Project on Artificial Intelligence* (DSRPAI)[11]. No início de 1920 o termo robô[12], foi cunhado por Karel Čapek e usado em sua peça *Rossumovi Univerzální Roboti*. Curiosamente, a peça também apresentou, pela primeira vez, robôs dominando o mundo, inaugurando um novo ramo da ficção científica.

Em 1939, o primeiro robô, ELEKTRO, foi exibido na feira mundial, marco na história da automação. Com comandos humanos, ele podia andar, fumar cigarros e explodir balões, i.e., um típico animador de festa infantil moderno. O sarcasmo aqui presente se faz para alertar o leitor sobre uma característica marcante atribuída a robôs, e que será herdada pela IA futuramente, o antropomorfismo.

Seguindo essa tendência, no início dos anos 40, Isaac Asimov criou as três leis da robótica. Adotadas e disseminadas por diversos autores em obras de ficção científica, as três regras (RL) determinam a maneira como o robô deve agir em relação aos seres humanos:

> (RL1) um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano seja prejudicado;
>
> (RL2) Um robô deve obedecer às ordens dadas pelos seres humanos, exceto quando essas ordens entrem em conflito com (RL1);
>
> (RL3) Um robô deve proteger sua própria existência, desde que essa proteção não entre em conflito com (RL1) e (RL2).

Em 1950, Alan Turing sugeriu uma maneira indireta de medir a "inteligência" de uma máquina, que testava a capacidade da máquina de "pensar", por meio de um jogo. O Teste de Turing, uma adaptação do jogo da imitação, ainda é usado como métrica de inteligência no campo da IA.

E por falar em jogo, em 1997, uma inteligência artificial chamada 'Deep Blue' derrotou Garry Kasparov no xadrez. 10 anos depois, foi a vez do jogo de damas ser decidido, i.e., o melhor resultado possível para um humano é empatar a partida. Em

---

[11] Vide (PICKOVER, 2019).
[12] Do tcheco, *robota* (i.e., trabalho forçado).

2016, o AlphaGo venceu, por 4-1, o Google DeepMind Challenge Match, i.e., campeonato de 5 partidas de Go[13], contra o campeão mundial Lee Sedol. Em 2017, O AlphaGoZero, uma versão do AlphaGo (que aprendeu a jogar Go treinando contra si mesma), demorou apenas 40 dias para bater o AlphaGo original.

Enquanto o Deep Blue usava chips VLSI personalizados para paralelizar o algoritmo de busca alfa-beta (GREENEMEIER, 2021), o AlphaGo e seus sucessores, suportados por milhares de CPUs e centenas de GPUs, usaram um algoritmo Monte Carlo de busca em árvore para selecionar seus movimentos com base em conhecimento previamente adquirido por aprendizado de máquina (SILVER, HUANG, *et al.*, 2016).

O AlphaGo usa técnicas de Deep Learning para treinar uma rede neural que identifica as melhores jogadas e os percentis vencedores dessas jogadas. A rede neural melhora a eficácia da busca em árvore, resultando em uma seleção de movimentos mais vantajosos a cada iteração (SILVER, HUANG, *et al.*, 2016). Já o Deep Blue, usava técnicas de *Good Old-Fashioned Artificial Intelligence* (GOFAI)[14] como minimax e a busca alfa-beta. O minimax é um algoritmo recursivo para escolher o próximo movimento em um jogo. A busca alfa-beta é um algoritmo de busca que visa diminuir o número de nós que são avaliados pelo algoritmo minimax em sua árvore de busca. O algoritmo de busca alfa-beta para de avaliar um movimento quando pelo menos uma possibilidade for encontrada que prove que o movimento é pior do que outro examinado anteriormente (HSU, CAMPBELL e HOANE, 1995). Tais movimentos não precisam ser mais avaliados. Quando aplicado a uma árvore minimax padrão, ele retorna o mesmo movimento que o minimax faria, mas elimina ramos que não podem influenciar a decisão final (Russell & Norvig, 2010).

Não obstante, por mais impressionante que seja a tecnologia, atual e futura, não devemos esquecer a história da automação nem o caminho percorrido para nos trazer onde estamos hoje. É impossível saber se os escritores de ficção científica estarão certos sobre uma futura rebelião da IA ou dos robôs; mas, se a tendepá em

---

[13] 圍棋 é um jogo de estratégia chinês, para dois jogadores, cujo objetivo é cercar mais territórios do que o oponente.

[14] GOFAI é um acrônimo cunhado por John Haugeland (HAUGELAND, JOHN, 1985) e refere-se apenas a um tipo restrito de IA simbólica (agentes lógicos ou baseados em regras), popular na década de 1980, especialmente empregada como uma abordagem para a implementação de sistemas especialistas (RUSSELL e NORVIG, 2021).

torno da inteligência artificial não é nova, o que exatamente é? Pelo menos duas diferenças caracterizam a conjuntura atual.

A primeira trata de fatores tecnológicos. Um grande aumento no poder computacional, na miniaturização e posterior massificação de dispositivos que coletam, geram, armazenam e compartilham, informações pessoais, aliados à massificação e pervasividade da Internet junto à população, saímos de 16% para 48% da população mundial com acesso à Internet entre 2005 e 2017 (ITU, 2015), nos permitiu alcançar a era do Big Data. Essa imensidão de dados sustentou o emprego das técnicas de inteligência artificial, em especial as de *machine learning* (ML), em diversos domínios, e.g., desde o diagnóstico de verrugas pré-cancerígenas até a condução de veículos, sobressaltando o potencial da IA para o bem e para o mal.

A segunda, dos formuladores de políticas públicas. Eles estão finalmente prestando mais atenção às consequências decorrente da massiva adoção da IA pelos diversos meios produtivos. A política admite a possibilidade de novas leis, mas não as exige. Pode até não ser sensato, ou mesmo viável, aprovar leis gerais sobre inteligência artificial nesta fase, principalmente quando apuramos os atuais e prováveis impactos sociais, econômicos, políticos e até ambientais da IA na sociedade hodierna e futura. Talvez mudanças menores nas normas, doutrinas e leis sejam mais apropriadas face às possibilidades positivas e negativas da IA.

Quiçá, as mudanças dimanem das práticas de governança adotadas pelas empresas de tecnologia. Diversas iniciativas como a Partnership on AI e a Open AI tem como objetivo estudar e formular boas práticas para tecnologias baseadas em inteligência artificial, permitindo sua evolução para um caminho que ajude a humanidade, e não o contrário. Infelizmente, como veremos a seguir, nem sempre prática e princípios andam em congruência.

1.2 FORMULAÇÃO DO PROBLEMA

Em The Black Box Society (PASQUALE, 2016), Frank Pasquale expõe como grandes corporações estão cocando nosso comportamento pessoal, i.e., examinando silenciosamente os vestígios deixados por nosso uso da Internet e nos incenti-

vando a consumir produtos, serviços e ideias. Os dados compilados e os perfis criados são incrivelmente detalhados, a ponto de serem invasivos.

Mas quem inspeciona o que as empresas estão fazendo com essas informações? Pasquale argumenta que todos precisamos ser capazes de fazê-lo, e exigir transparência é apenas o primeiro passo. Devemos estabelecer limites de como a big data afeta nossas vidas.

Uma sociedade inteligível garantiria que as principais decisões de suas empresas cardinais sejam justas, não-discriminatórias e abertas a críticas. O Vale do Silício e Wall Street precisam aceitar tanto *accountability* quanto impõem aos outros (PASQUALE, 2016).

Podemos definir *accountability* como uma norma: "A é *accountable* perante B quando A é obrigado a informar B sobre as ações e decisões de A (passadas ou futuras), justificá-las e sofrer punição em caso de má conduta eventual" (SCHEDLER, 1999). Todas essas obrigações e demandas supracitadas referem-se ao imo da responsabilidade algorítmica (*algorithmic accountability*).

Responsabilidade algorítmica refere-se à imputação de accountability sobre decisões (e ações) tomadas por pessoas (e sistemas) baseadas em deliberações algorítmicas. refere-se à atribuição de responsabilidade sobre como um algoritmo é criado e seu impacto na sociedade. E, em última análise, examina o processo de atribuição de responsabilidade por danos quando a tomada de decisão algorítmica resulta em efeitos discriminatórios e desiguais.

Para o nosso contexto, *algorithmic accountability* pode ser entendida como o processo de prestação de contas e prover explicabilidade sobre um determinado sistema ou IA. Ou seja, prestação de contas significa que, quem gera IA com relevância para a sociedade, e.g., as grandes empresas do Vale do Silício, deve regularmente explicar o que anda fazendo, como está fazendo, por qual motivo está fazendo, quanto gasta, o que está coletando, como está armazenando, sob quais bases legais opera etc.

Não se trata, portanto, de mera prestação de contas em termos quantitativos, mas, de autoavaliar a obra feita (e.g., produto, sistema, IA etc.), de dar publicidade aos resultados obtidos e de justificar aquilo que malogrou. A obrigação de prestar

contas, *lato sensu*, deve ser diretamente proporcional ao impacto sociopolítico que gera, e.g., COMPAS[15].

Explicabilidade pode ser definida como a característica do que é explicável, i.e., qualidade daquilo que se consegue explicar. A *General Data Protection Regulation* (GDPR), Regulamentação Geral de Proteção de Dados (RGPD) em português, garante o direito a uma explicação (i.e., explicabilidade). Para a GDPR, a concepção legal de explicações é tratada no artigo 5º, parágrafo 1, item h: "informações significativas sobre a lógica do processamento" (EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2016). Isso contrasta com o conceito de *black box*[16] em *machine learning*, onde nem mesmo os projetistas conseguem explicar por que a IA chegou a uma decisão específica (SAMPLE, 2017).

Felizmente, esse obscurantismo de certos modelos pode ser aclarado. A *Explainable Artificial Intelligence* (XAI), Inteligência Artificial Explicável (IAE) em português, refere-se a um rol de métodos e técnicas aplicados à inteligência artificial, de modo que as previsões geradas (i.e., resultados), o modelo e até os dados (usados para a criação do modelo) possam ser entendidos por seres humanos. Podemos então vislumbrar essas técnicas como um meio para prover o direito a explicação.

Contudo, se dispomos de XAI para deslindar as dinâmicas deliberativas da IA, nos resta obter os dados que alimentam, e embasam, esse esclarecimento. Quando um fato acontece (e.g., um acidente causado por um carro autônomo) todos os dados, análises, previsões, contenções e documentos que detemos, empregamos e planejamos sobre a IA até aquele momento são tidos como ex-ante, i.e., "antes do fato". Outrossim, toda e qualquer análise após o fato será ex-post.

Adotaremos a *Explainable Accountable Unit* (XAU), Unidade Responsiva Explicável em português (URE), como um coletor de dados, desenvolvido para atender as demandas específicas, que deve sempre registrar elementos informacionais sobre a IA que possam ser analisados e usitados para investigação de ocorrências, prestação de contas, depuração, imputabilidade etc.

---

[15] Vide item 4.5.1.2 COMPAS

[16] Dispositivo, sistema ou IA, que pode ser visualizado em termos de suas entradas e saídas sem nenhum conhecimento do seu funcionamento interno.

1.3 OBJETIVO DE PESQUISA

Esta tese tem como objetivos gerais: analisar, caracterizar e conformar *accountability* no âmbito da moral e da IA e prover um cânone que auxilie as diversas partes interessadas (e.g., gestores, políticos, juristas, ativistas, cidadãos, consumidores etc.) a lidarem com os anseios por responsabilidade algorítmica[17] (i.e., compreenderem, agirem, relatarem e – eventualmente – explicarem, consertarem e melhorarem os algoritmos que produzem, impactam ou que são afetados).

O arquétipo desenvolvido para exercer tais ofícios será o Metamodelo de *accountability* para Inteligência Artificial (MAIA), que ocorre durante o ciclo de vida da IA (i.e., visa governar a conformação dos processos, produção, cadeia de suprimentos etc.), e faculta as principais tarefas de *accountability*. Doravante, os objetivos específicos desta tese seguem.

Objetivo 1 – Definir *accountability* no contexto de Inteligência Artificial.

Apresentar, caracterizar e diferenciar conceitos de responsabilidade, responsabilidade civil, responsabilidade aquiliana e *accountability* na conjuntura sociopolítica normativa hodierna. E apresentar as duas exigências relacionadas a *accountability*.

Objetivo 2 – Apresentar o metamodelo de *accountability* para IA.

Que se respalda no metamodelo empregado pela Administração e Ciências Contábeis para prestação de contas de uma empresa, tanto na esfera pública como na privada e que consiste em quatro etapas e quatro questões normativas.

Objetivo 3 – Demonstrar a equivalência funcional do metamodelo contábil para IA.

Desenvolver a demonstração de equivalência entre o metamodelo contábil e o metamodelo de *accountability* para IA (MAIA) apresentando e embasando processos, técnicas, soluções e dificuldades.

---

[17] Conceito no qual as empresas devem ser responsabilizadas pelos resultados de seus algoritmos programados. Um termo correlato é a transparência algorítmica, que fomenta a predisposição das empresas em publicizarem propósito, estrutura e ações subjacentes dos algoritmos usados para pesquisar, processar e fornecer informações.

Há uma digressão quanto a comparação entre as técnicas de coleta e análise dos cômputos contábeis e informacionais (pois foge ao nosso escopo e seria deveras maçante) por isso, focamos apenas em apresentá-las para a IA quando abordamos XAI. Não há perdas, no entanto, pois a dimensão finalista de cada ferramental se manteve.

Objetivo 4 – Desenvolver cada uma das etapas e perguntas normativas do MAIA.

Sempre que possível, deslindar cada uma das quatro etapas e questões apresentando uma abordagem interpessoal e interpartes, principais técnicas, fontes, justificativas e processo e uma equivalência entre a aplicação do modelo contábil na escrituração da empresa e o MAIA na contabilidade da IA.

Objetivo 5 – Deslindar as técnicas de XAI e seu uso ex-ante e ex-post.

Apresentar algumas das principais técnicas de XAI, guardando suas características, origem, resultados, especificidades, escopo, compreensão e aplicabilidade. Classificar as referidas técnicas de acordo com sua origem, especificidade e origem. Ilustrar como seu uso ex-ante e ex-post pode dirimir questões refutáveis, políticas e normativas.

1.4 CONTRIBUIÇÕES

As contribuições do trabalho relatado nessa tese são resumidas a seguir.

(1) Definir *accountability* no contexto de Inteligência Artificial – i.e., *accountability* é ser *moralmente* responsável pelos sistemas de IA criados, operados, mantidos etc. e envolve dois deveres principais: realizar certas ações (ou abster-se de) e a prestação de contas pelas ações realizadas;
(2) Atestar a factibilidade de se inserir *accountability* no ciclo de vida da IA (capítulo 6);
(3) Adaptar o sistema contábil ao ciclo de vida da IA (capítulos 4 a 8);
(4) Metamodelo de *Accountability* para IA (MAIA) (capítulo 3);

(5) Comprovar a equivalência funcional entre o meta-arquétipo contábil e o MAIA (capítulos 3 a 8);
(6) Uso ex-post das técnicas de XAI como ferramentas forense (capítulo 8).

1.5 METODOLOGIA

Após uma extensa revisão da literatura acadêmica, técnica, governamental, jurídica e dos reguladores de políticas sobre o problema da explicação e auditabilidade da inteligência artificial (IA), sob uma perspectiva multidisciplinar, combinada com uma análise comparativa dos cenários europeu, brasileiro e norte-americano, o método empregado é essencialmente de natureza qualitativa. A tese e argumentos aqui expostos não serão sustentados por dados estatísticos nem alavancados em pesquisas de campo; ao contrário, surgiram da conceitualização e da reflexão sobre a natureza da noção de accountability, responsabilidade algorítmica, explicação, contabilidade e IA.

Ao longo dessa obra, optamos por diferenciar ética e moral (tratados de forma distinta no capítulo 4) pois não há um consenso, na literatura de Inteligência Artificial, a como se referir ao tema quando este é implantado numa IA. Evidente que para a maioria das escolas filosóficas ética e moral diferem apenas quanto etimologia da palavra (i.e., ética origem grega e moral origem latina) com ambas se referindo ao mesmo tema. Contudo, no corpo literário da área de IA moral e IA ética, há uma distinção entre os dois conceitos (SAVULESCU e MASLEN, 2015). Para a antologia citada, ética é uma teoria que se ocupa dos princípios que orientam as ações da IA, já a moral é prática e está relacionada às regras de conduta da IA. Por esse, e apenas por esse motivo, fizemos essa separação.

Também, devido a natureza multidisciplinar dessa tese, e seu enfoque em accountability e responsabilidade algorítmica, trataremos por IA toda a classe de algoritmos (e.g., busca e otimização, métodos probabilísticos para raciocínio incerto, classificadores e métodos estatísticos de aprendizado, redes neurais artificiais, lógicas e linguagens e hardware especializados) abarcados pelos diversos desígnios da matéria (e.g., raciocínio, resolução de problemas, representação do conhecimento,

aprendizagem, processamento de linguagem natural, percepção, inteligência social e inteligência geral) a menos quando explicitados. Isso se dá tanto por uma questão de simplicidade e legibilidade (evitando a enumeração desnecessária de técnicas, supra) quanto por finalidade, i.e., accountability (especialmente a obrigação de informar sobre ações e decisões e justificá-las) não se restringe a determinada definição ou tecnologia hodierna, passada ou futura de algoritmo. Portanto, circunscreve todos os algoritmos, independente da implementação, paradigma, finalidade ou complexidade.

Finalmente, resta-nos justificar por que a opção por um metamodelo. Como esta obra visa dirimir as principais questões pertinentes a *algorithmic accountability* – i.e., a obrigação de informar sobre como um algoritmo é criado e justificar seu impacto na sociedade – e prover um cânone para que os principais interessados lidem com os anseios por responsabilidade algorítmica – i.e., compreenderem, agirem, relatarem, explicarem, consertarem e melhorarem os algoritmos que produzem, impactam ou que são afetados – questionamo-nos sobre a factibilidade de um modelo ou processo para isso.

Num cenário em constante evolução e com uma pluralidade de técnicas de diferentes origens, complexidades, implementações e graus de inteligibilidade, prover explicações sobre o funcionamento desses algoritmos e justificar suas múltiplas repercussões sociais não parecia ser trabalho exequível para um simples processo ou modelo. Os processos mudam com o tempo, assim como os modelos de processo subjacentes a eles. Assim, novos processos e modelos podem ter que ser construídos e os existentes melhorados para se adequarem a novas demandas (ROLLAND, 1998).

Até poderíamos construir modelos específicos para as principais técnicas empregadas atualmente na área de Inteligência Artificial, mas seria deveras maçante e rapidamente ficariam datados, quer pela evolução da técnica quer pelo surgimento ou adequação a novas demandas informacionais e sociais. Fez-se necessário portanto o emprego de um ferramental que gerasse esses modelos e processos a luz das novas demandas, i.e., metamodelos e metaprocessos.

Um metamodelo ou modelo substituto é um arquétipo do modelo, i.e., uma representação simplificada de um modelo real e.g., um circuito, sistema ou software

como entidade. Metamodelagem é a análise, construção e desenvolvimento de estruturas, regras, restrições, modelos e teorias aplicáveis e úteis para modelar uma classe pré-definida de problemas (GARITSELOV, MOHANTY e KOUGIANOS, 2012).

A modelagem de metaprocesso se concentra e dá suporte à técnica de construção de metamodelo de processos. A sua principal preocupação é melhorar os modelos de processos e fazê-los evoluir, o que por sua vez servirá de suporte ao desenvolvimento de sistemas (ROLLAND, 1998). Um modelo de processo é uma instanciação de um metamodelo de processo que por sua vez, pode ser instanciado várias vezes para definir vários modelos de processo (ROLLAND, 1998).

Munidos desses novos aprestos (metamodelo, metaprocesso e metamodelagem), erigimos o nosso metamodelo (MAIA) fundamentado no já consolidado e verificado metamodelo contábil. A escolha pelo metamodelo contábil se deu não só pelo seu historial e validação, que nosso metamodelo herda, mas por similitude de finalidade (coletar, analisar, conformar e divulgar informações). Assim, nosso metamodelo nasce fausto, pois goza de validação sempiterna, cabendo-nos construir durante essa produção, por meio de modelagem de metaprocesso, o mapeamento entre os processos do metamodelo contábil e os processos do MAIA.

Por exemplo, se o metamodelo contábil diz que devemos nos preocupar onde, quando e como coletar informações contábeis, o MAIA também dirá que devemos nos preocupar onde, quando e como coletar informações sobe a IA e, assim como os metaprocessos contábeis nos mostram como gerar modelos que viabilizam essas demandas, também os temos no MAIA. Por último, ressaltamos que nosso mapeamento e metamodelagem seguiram de forma ad hoc para que fosse o mais acessível possível, sempre tivemos em mente aclarar e incluir o maior espectro possível (usuários, juristas, formadores e tomadores de opinião etc.) em nossa audiência.

1.6 ESTRUTURA DA TESE

Esta tese consiste em nove capítulos e um anexo. O primeiro capítulo oferece uma introdução, onde são fornecidas a motivação (1.1 M), a formulação do problema (1.2 Formulação do problema), os objetivos (1.3 Objetivo de pesquisa), as contribuições (1.4 Contribuições), a metodologia (1.5 Metodologia) e a estrutura da tese (1.6 Estrutura da tese).

O segundo capítulo aborda e caracteriza o conceito de *accountability*. São apresentados, caracterizados e diferenciados conceitos de responsabilidade, responsabilidade civil, responsabilidade aquiliana e *accountability* na conjuntura sociopolítica normativa hodierna. Nele são apresentadas as duas exigências relacionadas a *accountability* e conceitos como *accounting*, *data analytics*, decisão informada e as seis características da informação de qualidade.

O terceiro capítulo apresenta o Metamodelo de *Accountability*. Começamos por introduzir os sistemas de *accountability* empresarial (3.1 Sistemas de *Accountability* Empresarial) e informação contábil (3.1.1 Sistema de Informação Contábil (SIC)). Em seguida, são veiculadas as quatro questões normativas que compõem o metamodelo (“3.2.1 (QN1) Por que coletar e divulgar informações?”, “3.2.2 (QN2) Quem são os stakeholders contemplados por essas informações?”, “3.2.3 (QN3) Quais tipos de informações serão coletadas e quais divulgações serão feitas?” e “3.2.4 (QN4) Como as informações devem ser divulgadas?”), suas respectivas justificativa e finalidade, além dos conceitos de escopo (3.3 Escopo) e recursos (3.4 Recursos). No tópico (3.5 MAIA) é conformado o MAIA.

O quarto capítulo apresenta a primeira questão normativa (QN1 - Por que coletar e divulgar informações?). Nele são dissertados os principais fatores que regulam a divulgação de informações sobre a IA: ética (4.1 Ética), moral (4.2 Moral), reputação (4.3 Reputação), normas (

4.4 Normas), leis (4.5 Leis), políticas (4.6 Políticas) e educação (4.7 Educação). Cada um desses fatores é tratado *per se* e, correferido com a IA.

O quinto capítulo desenvolve a segunda questão normativa (QN2 - Quem são os stakeholders contemplados por essas informações?). Lá são arrazoados os temas da identificação (5.1 Identificação dos stakeholders), categorização (5.2 Categorização dos stakeholders), priorização (5.3 Priorização dos stakeholders), mapeamento (5.4 Técnicas de mapeamento de partes interessadas) e comunicação (5.5 Comunicação com os stakeholders) das partes interessadas.

O sexto capítulo oferece um parêntese no alude as questões normativas, e nos brinda com uma visão holística sobre quando, onde e como coletar dados sobre a IA. O capítulo retrata o momento da coleta (ex-ante ou ex-post) tópico 6.1 Quando coletar, o local da coleta (ciclo de vida da IA, aqui exemplificado pelos processos KDD e CRISP-DM)[18], item 6.2 Onde coletar, e a forma (sistemas de telemetria (6.3.1 Telemetria), aqui ilustrada pela *Explainable Accountable Unit* (XAU), ponto 6.3.1.1.2.1 Explainable Accountable Unit (XAU).

O sétimo capítulo retoma a retratação das nossas questões normativas com a apreciação da (QN3 - Quais tipos de informações serão coletados e quais divulgações serão feitas?). Começaremos por definir o escopo, cômputos internos e externos, (i.e., 'quais divulgações serão feitas') e recursos, o que, onde e como medir, (i.e., 'quais tipos de informação serão coletados'). Reservamos o terceiro item 7.3 XAI, para expor o conceito de explicabilidade (7.3.1 Explicabilidade), uma taxonomia dos seus métodos (7.3.1.1 Taxonomia dos Métodos de Explicabilidade), algumas de suas principais técnicas (7.3.1.2 Técnicas de Explicabilidade) e um subconjunto temático, sobre imparcialidade (7.3.2 Imparcialidade). Essas técnicas constituem um dos eixos que respaldam o emprego e necessidade da (XAU).

O oitavo capítulo discute a quarta, e última, questão normativa (QN4 - Como as informações devem ser divulgadas?). Começaremos por compreender como se dá o processo de apuração dos dados até se tornarem conhecimento. Ulteriormente, examinaremos no item 8.2 Análise de Acidente a técnica de Análise de Acidente, e suas quatro etapas. Em seguida, no tópico 8.3 Análise da Causa Raiz (ACR), veremos as sete etapas da Análise da Causa Raiz (ACR) e como ela, juntamente com a Forense Digital (FD), item 8.4 Forense Digital, promoverão a preservação, coleta, validação, identificação, análise, interpretação, documentação e apresentação de

[18] CRoss-Industry Standard Process Model (CRISP-DM) e Knowledge Discovery in Databases (KDD)

evidências digitais derivadas de fontes digitais usadas para identificar as causas raiz de falhas ou defeitos.

Por fim, no ponto 8.5, examinaremos o uso da XAI como técnica forense, i.e., Inteligência Artificial Explicativa Forense (IAEF)[19]. Ademais, explanaremos sobre as principais fases que diferem das nove etapas do processamento de evidências digitais nos tópicos: 8.5.1 Exame, 8.5.2 Análise e 8.5.3 Apresentação.

O nono capítulo expõe considerações finais, contribuições da pesquisa e trabalhos futuros. Cabe ao Anexo I examinar as principais técnicas de explicabilidade de IA. São dissertados dois grupos: explicabilidade, i.e., técnicas que visam aclarar a mecânica da IA, e neutralidade, i.e., técnicas que ambicionam dirimir vieses e injustiças nos algoritmos.

[19] Vide (PADOVAN, MARTINS e REED, 2022)

2 ACCOUNTABILITY

Neste capítulo, nos debruçaremos sobre as principais definições que pesam para a definição de *accountability*. Apesar dos primeiros indícios, as diversas partes envolvidas terão compreensões distintas, e quase sempre conflituosas, sobre o que é *accountability* e o que abarca (i.e., atribuições e deveres).

Serão apresentados, caracterizados e diferenciados conceitos de responsabilidade na conjuntura sociopolítica normativa hodierna findando com as duas exigências relacionadas a *accountability*. Posteriormente, em 2.1 *Accounting*, trataremos de conceitos como *accounting*, *data analytics*, decisão informada e as seis características da informação de qualidade.

Antes de adentrarmos na tarefa de definir *accountability* propriamente, vamos tratar de outros dois termos que aparecem sempre que discutimos o tema: responsabilidade (*responsibility*) e responsabilidade civil (*liability*).

> **Responsabilidade** - demonstra a qualidade de quem é responsável, ou a obrigação moral, jurídica ou profissional de responder por atos próprios, alheios, ou por coisa confiada relacionados ao cumprimento de determinadas leis, atribuições ou funções.

Da definição acima sobressaem três características essenciais para distinguirmos *liability* e *accountability*: domínio, agente e normas. Domínio é o campo de ação ou extensão de poder e influência (e.g. moral, jurídica ou profissional), agente é aquele que exerce uma ação que produz algum efeito e normas são diretrizes que estabelecem e regulam procedimentos (e.g. leis, atribuições ou funções).

> **Responsabilidade civil** - é toda ação ou omissão que gera violação de uma norma jurídica e causa prejuízo a outrem. Assim, nasce uma obrigação de reparar o ato danoso.

Tanto domínio quanto normas são evidentes na definição acima. O elemento novo que surge é o prescritivo (i.e., o dever). Como veremos adiante, a classe das normas que aplicamos em nosso modelo de responsabilidade impacta diretamente nos direitos, deveres e obrigações do agente.

Na sua maioria, as normas legais são injuntivas, i.e., sua aplicação é independente da vontade dos destinatários, estabelecendo-se uma consequência jurídica (RENO, CIALDINI e KALLGREN, 1993). Podem ser proibitivas como preceptivas. As normas proibitivas, vedam determinadas condutas ou comportamentos. Já as normas preceptivas impõem determinado comportamento ou conduta, a realização de determinada ação.

É possível também que empreguemos normas de um domínio para analisar outro. Esse tipo de análise é útil, especialmente quando queremos avaliar riscos ou a multiplicidade de consequências de um determinado ato nos diversos domínios que tange. Como exemplo, temos a responsabilidade aquiliana.

> **Responsabilidade aquiliana** - responsabilidade extracontratual. É a responsabilidade que decorre da inobservância de norma jurídica, por aquele que, por ação ou omissão voluntária, negligência ou imprudência, viola direito e causa dano a outrem, ainda que exclusivamente moral (ACQUAVIVA, 2001).

Vale evidenciar que, por vezes, um determinado fato, quando abordado em diferentes domínios e normas apropriadas, pode apresentar resultados variegados. Por vezes, isso se dá devido a mudança da natureza da norma que versa sobre o assunto nos diferentes domínios, e.g., normas legais injuntivas e as morais dispositivas, i.e., só se aplicam se as partes suscitarem, ou não afastarem, a sua aplicação.

Outras vezes, o contexto sociopolítico será preponderante para a divergência, mesmo quando comparamos domínios iguais. Por exemplo, segundo o código de ética médica, ao médico é vedado abreviar a vida do paciente, ainda que a pedido deste ou de seu representante legal (CONSELHO FEDERAL DE MEDICINA, 2019). O suicídio assistido e a eutanásia são proibidos no Brasil, artigo 122 e artigos 121 e 29 do Código Penal respectivamente. Atualmente, a eutanásia ou suicídio assistido por médico pode ser praticado legalmente na Holanda, Bélgica, Luxemburgo, Colômbia e Canadá (Quebec desde 2014, nacionalmente a partir de junho de 2016). O suicídio assistido por médico, excluindo a eutanásia, é legal em 5 estados dos EUA (Oregon, Washington, Montana, Vermont e Califórnia) e na Suíça (EMANUEL, ONWUTEAKA-PHILIPSEN, *et al.*, 2016).

Após discernirmos as três características essenciais da responsabilidade, as singularidades que guardam cada par (domínio, norma) e aludirmos sobre divergências cabíveis quanto a mudança de contexto, nos cabe ainda a tarefa de definir *accountability*. Isto é, quais são o domínio, agente e normas da *accountability*?

Accountability é o dever de prover avaliação ou cômputo das ações pelas quais alguém é considerado responsável (GRAY, ADAMS e OWEN, 2014). Envolve a obrigação de responder pelas decisões e ações de alguém quando a autoridade para agir em nome de uma parte (o principal) é transferida para outra (o agente) ... *Accountability* requer abertura, transparência e o fornecimento de informações, e a aceitação de responsabilidade pelas próprias ações (BARTON, 2006).

Accountability é a obrigação de um indivíduo ou organização de responder por suas atividades, aceitar a responsabilidade por elas e divulgar os resultados de maneira transparente (BUSINESSDICTIONARY.COM, 2017). Assim, dizemos que A é *accountable* perante B quando A é obrigado a informar B sobre as ações e decisões de A (passadas ou futuras), justificá-las e sofrer punição em caso de má conduta eventual" (SCHEDLER, 1999). Ser *accountable* é ser responsável por ações e decisões e explicá-las quando solicitado. É ser oficialmente responsável perante o público, de modo que a ênfase é colocada sobre a posição que você ocupa e nos deveres relacionados a ela.

Todas essas definições arrolam uma série de cuidados, deveres e características anunciando *accountability* como uma classe de responsabilidade cujo domínio é a moral e/ou profissional. As normas são primordialmente reputacionais, cabendo ainda leis, atribuições e funções. Ademais, sinalizam que *accountability* resulta de um momento anterior e quase sempre negligenciado, o pacto.

Se pensarmos *accountability* como o cumprimento de um pacto entre duas pessoas, podemos defini-la simplesmente como a capacidade de um indivíduo satisfazer as expectativas de um terceiro, i.e., A é responsável por fazer algo para B.

Ao tomarmos a definição de *accountability* como a efetivação de uma promessa, fica fácil entender por que um indivíduo deve responder, explicar, aceitar responsabilidades e eventuais punições por suas ações, i.e., tarefas inerentes ao cumprimento do pacto. Ademais fica evidente o fato de que *accountability,* por se

tratar de um relacionamento, requer como agentes um solicitante (demandante) e um concedente (demandado).

Na sua manifestação mais simples, solicitante e concedente são pessoas naturais e o processo de *accountability* é iniciado pelo solicitante. A concedente, caso aceite a proposta, se compromete com o solicitante e cumpre sua promessa. No entanto, é fácil perceber que nem sempre esse processo se dará entre pessoas naturais (e.g., fornecedores, governo, credores), a ação nem sempre é iniciada pelo solicitante (e.g., o comprometimento pode partir do próprio demandado), nem todo comprometimento é uma promessa (e.g., carta de intensão) e nem sempre podemos cumprir com nossas promessas (e.g., por motivos alheios ou não a nossa vontade como a morte de um colaborador ou falha no planejamento). Assim, é fácil perceber por que ser *accountable* perante alguém exige informá-la sobre suas ações e decisões, justificá-las e sofrer punição.

Em última análise, o pacto sempre será praticado entre pessoas naturais. Com isso, queremos dizer que mesmo os acordos firmados entre empresas e governo, por exemplo, serão consumados por pessoas naturais (colaboradores internos e externos a instituição a depender do pactuado) e, geralmente, envolvendo vários cooperários, em uma rede, para esse fim.

Por exemplo, uma empresa que se compromete a adotar práticas mais sustentáveis, nesse caso teríamos uma pessoa jurídica (empresa) se comprometendo com a sociedade, consumidores, fornecedores, credores e demais parceiros (implícitos), terá seu presidente e demais funcionários formando uma rede (concedente) que busca realizar a adoção das práticas sustentáveis (pacto). Os interessados, mesmo implícitos, serão pessoas físicas e jurídicas interessadas e impactadas por essas ações. Essas partes terão necessidades informacionais sobre o andamento, eficácia, custos etc. bem como exigiram justificativas e punições em casos de inconformidades.

Especificamente sobre as punições, lembremos que estamos no domínio da moral e, portanto, se darão, principalmente, sobre a reputação da empresa. O problema nesses casos é que por se tratar de uma pessoa jurídica, a “culpa” pode ser entendida como da empresa, salvaguardando as pessoas, ou de parte da rede colaborativa para a efetividade do pacto.

Distinguir entre *accountability* pessoal e organizacional nem sempre é tarefa fácil. Sem conhecer a cultura da empresa, seus processos, organogramas e *accountability* organizacional, no que toca especialmente culpabilidade, nos vemos tentados a imputar responsabilidade sobre o encarregado imediato ou sobre o gerente da seção responsável pelo pacto. Tratamos aqui da reputação (e valores como honestidade, justiça, integridade) da empresa ou de seus funcionários, o que não exclui punições legais cabíveis.

Uma análise (com escopo alargado que contemple questões morais, profissionais, sociais, econômicas e jurídicas) é essencial antes de firmarmos e para cumprirmos o tratado. Geralmente, concluímos que algumas das nossas promessas deveriam ser intenções, e.g., algo que pretendemos fazer (depositamos e criamos expectativas) e que não há garantia de realização.

Faz-se necessário portanto que ambas as partes, i.e., solicitante e concedente, sejam ativos durante o processo de pactuação. Do solicitante, ex-ante, espera-se um engajamento no processo e que seja claro quanto ao que deseja e da concedente, que observe as demandas do solicitante e que seja clara quanto às suas possibilidades de cumpri-las. Durante o pacto, o solicitante deve acompanhar o andamento e cumprimento, reagindo quando necessário enquanto a concedente deve comprometer-se, a informar e assumir eventuais problemas. Ex-post, as partes devem avaliar a efetiva realização do pacto e a concedente aprender com os erros e corrigir eventuais falhas.

Doravante, nos concentraremos no relacionamento entre empresas produtoras de sistemas de IA (concedentes) e seus solicitantes. Não excluímos a possibilidade de uma pessoa natural ser a única responsável por um sistema, mas essa é a insignificância dos casos. À vista disso, *accountability* pode ser entendida como o estado de ser responsável e prestar contas sobre um determinado sistema ou IA.

Essa responsabilidade se assemelha a noção de responsabilidade civil decorrente de ato de terceiros (*vicarious liability*) que decorre de ato de terceiro, com o qual o agente tem vínculo legal de responsabilidade. Ou seja, *accountability* é ser moralmente responsável pelos sistemas de IA criados, operados, mantidos etc.

Accountability – envolve duas responsabilidades ou deveres principais:

1. Realizar certas ações, ou abster-se de, de acordo com as expectativas de um grupo (solicitantes).
2. Fornecer uma avaliação, ou cômputo, dessas ações para os interessados.

A definição acima de *accountability* implica uma medida de exigibilidade "*enforceability*"(realizar certas ações, ou abster-se de) e responsividade "*answerability*" (prestação de contas pelas ações realizadas) (GOETZ e JENKINS, 2001).

A natureza das avaliações, cômputos ou relatórios produzidos por uma organização, e os indivíduos ou grupos interessados que terão acesso a estes pareceres, serão influenciados por julgamentos relativos às responsabilidades da organização e suas incumbências associadas. Isso porque existe uma relação direta entre as responsabilidades (ou deveres) que a gerência da organização acredita ter e os compromissos aceitos. Isto é, se a superintendência da organização se compromete junto a uma determinada parte por certos aspectos da IA, e.g., imparcialidade, ela também aceita que essa parte tenha o direito de obter informações sobre esses aspectos.

A questão agora é estabelecer como construir estes cômputos dadas as obrigações e expectativas dos interessados. Para tal, as empresas lançam mão do processo de *accounting*.

2.1 ACCOUNTING

Accounting (contabilidade em português) – é simplesmente o processo de identificar, medir e relatar informações econômicas sobre uma organização, para permitir que decisões embasadas sejam tomadas por indivíduos, ou grupos, com participação financeira nessa organização (GRIFFIN, 2012). Pode ser administrada por um contador, e.g., em uma pequena empresa, ou pelo departamento financeiro, com dezenas de funcionários, em empresas de grande porte.

Mas o negócio não sobrevive isolado. O contexto no qual ele está inserido e opera refletem direta e indiretamente na sua saúde financeira e, por conseguinte, no

seu sucesso. À vista disso, segundo a máxima atribuída a William Edwards Deming, “O que não se mede, não se gerencia” e, portanto, devem fazer parte da contabilidade não só fatores financeiros, como outros que impactem na supervivência da companhia.

Ao dilatarmos o escopo da contabilidade para além do âmbito financeiro, e.g., plano socioambiental, podemos nos valer de fluxos semelhantes para identificar, medir e relatar informações socioambientais sobre a organização. Portanto, uma visão ampliada deve capturar os vários aspectos do desempenho da empresa, e.g., financeiro, social e ambiental.

> **Contabilidade** – é uma prática técnica e social. Pode ser definida como o processo de coleta, resumo, análise e comunicação de informações para permitir que os usuários dessas informações tomem decisões informadas.

Mas, como contabilidade e IA se relacionam? Para respondermos essa interpelação, é interessante notar a semelhança entre a definição acima de contabilidade e a definição de *data analytics*.

> ***Data analytics*** (análise de dados em português) – é a ciência de integrar dados heterogêneos de diversas fontes, fazer inferências e previsões para permitir a inovação, obter vantagem competitiva nos negócios e ajudar na tomada de decisões estratégicas (CHOWDHURY, APON e DEY, 2017). O domínio de análise de dados evoluiu sob vários nomes, incluindo *online analytical processing* (OLAP), mineração de dados, análise visual, análise de big data e análise cognitiva.

O fulcro aqui é que o termo *analytics*, i.e., análise, é usado para se referir a qualquer tomada de decisão baseada em dados, i.e., independente da natureza ou finalidade desses dados. Em síntese, contabilidade e *data analytics* tratam de coletar, analisar e comunicar dados para permitir que os usuários desses elementos tomem decisões informadas.

Doravante, tomaremos o termo análise factual para referirmo-nos ao metaprocesso de coletar, analisar e comunicar informações para permitir que os usuários dessas informações tomem decisões informadas. Mas o que são decisões informadas?

Decisão informada (decisão fundamentada, embasada, factual) – É a decisão baseada em informação fiável (baseada em fatos, e.g., dados) e relevante (e.g., compreensível, verificável).

A análise factual fornece a várias pessoas, dentro e fora de uma organização, informações para tomada de decisão, e é papel do curador (e.g., contador ou administrador de dados) determinar quais informações são mais apropriadas para permitir que essas pessoas tomem decisões informadas. Mas o que informar? E para quem informar?

Usualmente informamos o desempenho, i.e., resultados, impactos ou realizações, associados à conclusão de uma determinada tarefa ou grupo de tarefas. Na contabilidade são três fluxos: performance financeira, social e ambiental.

A performance financeira consiste em uma medida ou avaliação do desempenho de uma organização em termos financeiros, geralmente através do uso de padrões de contabilidade financeira. A performance social visa apreciar os impactos, positivos e negativos, que as atividades de uma organização têm sobre suas partes interessadas, e.g., funcionários, clientes e a comunidade em geral. A performance ambiental pesa os impactos, positivos e negativos, que a organização tem nos ambientes físicos e naturais em que opera.

No campo da IA, usualmente informamos o desempenho, i.e., resultados, associados à conclusão de uma determinada tarefa ou grupo de tarefas, e.g., reconhecimento, planejamento ou predição. No modelo de inteligência artificial explicável (*explainable artificial intelligence*), são três os principais fluxos: explicabilidade, imparcialidade e resiliência.

Explicabilidade visa aclarar a mecânica da IA. Consiste em um conjunto de medidas ou avaliações de desempenho da IA de acordo com vários critérios, e.g., origem, resultado, especificidade e escopo. Os métodos para apurar imparcialidade geralmente analisam as previsões do modelo para segmentos demográficos sensíveis (e.g., etnia, gênero, cor, sexualidade), mas podem abordar casos individuais, i.e., verificar a inconformidade de uma previsão específica. Os métodos que avaliam resiliência mensuram a capacidade da IA em lidar com problemas, adaptar-se a mu-

danças, superar obstáculos ou resistir a ataques que possam sofrer em ambientes físicos e naturais onde opera.

Um elo entre as percepções de responsabilidade, ou dever, e a prestação de contas vincula as atribuições de uma organização à obrigação de fornecer informações sobre seu desempenho e sua aderência às expectativas dos outros. Contudo, ainda nos resta responder a quem devemos informar.

> **Stakeholders** (usuários ou partes interessadas) – são qualquer grupo ou indivíduo identificável que pode afetar a realização dos objetivos de uma organização, ou é afetado pela realização dos objetivos desta (FREEMAN e REED, 1983).

Entendimento análogo pode ser empregado no caso da IA. A fim de ilustração, são exemplos de grupo ou indivíduo identificável que pode afetar a realização dos objetivos da IA: o administrador de dados e os demais encarregados da gênese e manutenção da IA, a gerência da empresa, legisladores e membros do judiciário, usuários, terceiros etc. Como exemplo de afetado pela IA, temos os usuários como epítome. Quais stakeholders devem receber esclarecimentos sobre quais performances, e quais aspectos dessa performance devem ser informados depende de cada caso, i.e., do que foi pactuado.

A lei impõe certos requisitos mínimos em termos dos cômputos que devem ser criados (e.g., a avaliação de impacto demandada pela AAA22)[20]. Se houver regulamentação que exija especificamente que um determinado grupo (ou grupos) de stakeholders receba informações específicas, espera-se que essas informações sejam divulgadas.

Por exemplo, para a *General Data Protection Regulation*, a concepção legal de explicações é tratada no artigo 5º, parágrafo 1, item h: “informações significativas sobre a lógica do processamento” (EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2016). Destarte, informar quais os principais critérios são relevantes para uma análise de crédito realizado por uma IA ou afigurar contraexemplos para ilustrar uma recusa, integram aspectos relevantes.

---

[20] Vide item 4.6 Políticas

As organizações, normalmente, fornecem muitas outras análises de forma voluntária, e a natureza desses balanços será influenciada pelas responsabilidades que a administração acredita e aceita. Todavia, assuntos que não são regulamentados são divulgados a critério da gerência. Retomando o exemplo anterior, informar a performance de resiliência da IA aprovadora de crédito para o público não parece coerente, mas a performance de imparcialidade da IA, mesmo que não obrigatória, deve ser considerada sob a luz da transparência e equidade.

Ponderaremos agora sobre algumas das características dos cômputos gerados pelo processo de análise factual. Refletiremos sobre o escopo, quem produz, e sobre os critérios de qualidade dessas demonstrações.

Sabemos que as informações coletadas para fins de gerenciamento interno podem ou não ser fornecidas a pessoas externas à empresa. Isto posto, referente ao escopo, temos dois grupamentos de informações: cômputos internos e cômputos externos (ALBRECHT, STICE, *et al.*, 2008).

> **Cômputos internos** – são informações sobre a posição e o desempenho, da empresa ou da IA, destinadas ao uso por pessoal interno a organização, e.g., gerentes, diretores, funcionários. Para ilustrar, temos a folha de ponto ou a escala de funcionários, organogramas, e no caso da IA, explicações sobre seu funcionamento, sua arquitetura, imparcialidade, resiliência.
>
> **Cômputos externos** – são informações sobre a posição e o desempenho, da empresa ou da IA, destinadas ao uso por pessoas externa a organização, e.g., acionistas e reguladores. Para ilustrar, temos o balanço patrimonial e o relatório de impactos socioambiental da empresa, e no caso da IA, explicações sobre seu funcionamento ou métricas de desempenho.

Aqui, cabe ressair que foi deliberado ilustrar, no caso da IA, “fornecer explicações sobre seu funcionamento”. Não só por se tratar de uma obrigação legal, como explanado anteriormente, mas para aludir três outros aspectos atinentes ao escopo, apresentados sob as seguintes indagações: que informações devem ser coletadas? quanta informação deve ser divulgada? Para quem revelar essas informações?

As informações coletadas estão intimamente vinculadas a quais aspectos você acredita que a organização, ou a IA, é responsável e, portanto, deve esclareci-

mentos. Isso pode exigir anteferir os impactos em termos de quais criam os maiores efeitos negativos ou positivos.

Presumivelmente, priorizamos a divulgação de informações valiosas para nossos stakeholders. Isto é, as informações divulgadas estão associadas a quais aspectos da organização, ou da IA, eles têm maior interesse.

Na prática, como a divulgação de muitos itens de informação é voluntária (principalmente em relação aos aspectos sociais) alguns gestores tendem a divulgar informações de natureza positiva ou favorável, em vez de informações de natureza negativa. Portanto, as partes interessadas precisam considerar a parcialidade e a incompletude das informações divulgadas voluntariamente.

Quanto ao agente gerador, a priori, qualquer pessoa dentro da organização gera informação sobre vários processos e, portanto, contribui para a análise factual. Tal-qualmente, não são só as escolhas do administrador de dados que impactam na IA, mas os demais colaboradores, e.g., diretores e funcionários correlatos, também. Para explanar tomemos a imparcialidade. Se a gerência não endossar tal aspecto, dificilmente haverá averiguação por parte da equipe de desenvolvimento. Da mesma forma que se não houver engajamento dos demais colaboradores afetados pela IA, provavelmente não teremos a pluralidade de casos a serem observados, o que acarreta numa avaliação deficitária.

Retomemos a finalidade da análise factual, i.e., geração de cômputos detalhando a performance sobre um determinado aspecto. As informações devem ser produzidas para que os vários interessados tomem decisões informadas sobre como irão gerenciar a organização (no caso de gerentes) ou se irão apoiá-la (externos) em termos de atividades como investimento, emprego, empréstimo de fundos ou consumo de bens ou serviços. Um bom exemplo na IA seria a adoção de um determinado reconhecedor facial. De posse tão somente de métricas gerais de precisão e recall não temos como avaliar o comportamento dela ante um ataque de evasão, e.g., usar uma máscara ou maquiagem para enganá-la, ou sua precisão em reconhecer as diversidades étnicas da população brasileira.

Mas, quais características deve possuir uma informação para ser dita de qualidade? Não há um consenso na literatura sobre quais atributos definem a qualidade

de uma informação, mas as seguintes características qualitativas devem estar presentes para serem úteis na tomada de decisões: relevância, confiabilidade, comparabilidade, verificabilidade, compreensibilidade e disponibilidade.

> **Relevância** - a informação é relevante se for capaz de mudar as decisões das pessoas que a recebem.

Determinar quais informações são relevantes depende de um alto grau de julgamento profissional e conhecimento tácito. Sua relevância geralmente é afetada pela localização da organização (e.g., diversidades étnicas e culturais), bem como, pelo tempo (e.g., validade das informações), i.e., princípios da localidade e temporaneidade. As informações relevantes terão valor preditivo ou valor confirmatório. A informação tem valor preditivo se permite a previsão de eventos futuros, ao passo que tem valor confirmatório se nos permite confirmar eventos passados.

> **Confiabilidade** - a informação é confiável se estiver livre de erros, completa e imparcial.

Embora haja uma expectativa de que as informações sejam confiáveis, isso não significa que se deva sempre esperar que as informações sejam perfeitamente precisas, pois na prática isso raramente é possível. Isso vale para sua completude e imparcialidade, que pode se demonstrar impossível. Significa simplesmente que se deve tomar cuidado para tornar a informação a mais confiável possível. Esse aspecto geralmente é trabalhado nas etapas de *data quality* e *data preparation* no processo de mineração de dados.

Relevância e confiabilidade são normalmente consideradas as características qualitativas mais importantes, ou fundamentais, que as informações devem possuir. Contudo, outras quatro são essenciais para a atingirmos nosso objetivo de prover *accountability*.

> **Comparabilidade** - a informação é comparável se selecionada, compilada e medida de forma semelhante, de forma que relatórios de diferentes períodos (e diferentes instâncias) possam ser comparados (benchmarking). Por exemplo, podemos querer comparar a evolução do nosso classificador de imagens entre suas diferentes versões ou compará-lo com outros classificadores.

Para auxiliar na obtenção da comparabilidade, é útil para os preparadores da informação fornecerem notas claras, geralmente metadados, sobre os métodos e políticas usadas para mensurar itens específicos. Retomando o exemplo anterior, sem saber, e justificar, os métodos e políticas empregados na análise dos classificadores, nos permite sempre alegar que um classificador é imparcial ou confeccionar uma base de comparação favorável ao nosso classificador. Se diferentes métodos foram usados para medir um item específico, então sabemos que se deve tomar cuidado ao comparar as diferentes medições.

Verificabilidade - em termos de informação, quando métodos semelhantes são empregados, espera-se resultados semelhantes.

Se pessoas diferentes receberem os mesmos dados subjacentes, i.e., dados estatisticamente próximos aos que utilizamos para gerar os demonstrativos, elas derivarão os mesmos resultados (i.e., com baixa divergência) mostrados nas informações apresentadas pela organização. Seguindo com o exemplo da IA classificadora, é de se esperar que ao não se levar em conta a etnia da pessoa na rotulagem, pessoas de etnias diferentes possuam índices próximos de acerto, e erro. Raciocínio análogo para outras características como cor dos olhos ou presença de pelos faciais.

Compreensibilidade - em termos de informação, o seu significado e a base da sua compilação são claros para o utilizador, i.e., os utilizadores compreendem o que significa a informação e a base em que foi medida.

O conhecimento dos stakeholders deve ser considerado ao fornecer-lhes informações. Imagine que você desenvolveu uma IA que estima a chance de um paciente desenvolver cardiopatia baseado, entre outros fatores, na altura e no peso do paciente. Provavelmente, se o seu sistema foi desenvolvido no sistema métrico, espera-se valores entre 1 e 2 para a altura e entre 40 e 150 para o peso (claro que valores fora desses limites existem, mas não há registro de ninguém com 3 metros[21] ou mais de altura ou 1 tonelada[22] de peso). No entanto, esse mesmo sistema sendo

[21] O recorde de homem mais alto, já registrado, é do americano Robert Wadlow, 272cm (GUINNESS WORLD RECORDS LIMITED, 1955)
[22] O recorde de homem mais pesado, já registrado, é do americano Jon Brower Minnoch, 635Kg (GUINNESS WORLD RECORDS LIMITED, 1978)

utilizado no sistema imperial receberia valores em pés e libras, resultando em uma margem diferente. Esse aspecto geralmente é trabalhado nas etapas de *data understanding* e *data preparation* no processo de data mining.

> **Disponibilidade** - em termos de informação, os usuários devem ter acesso a ela, a tempo de tomar decisões factuais. Geralmente, quanto mais antiga a informação, menos útil ela é.

Não estamos, contudo, dizendo que informações antigas são inúteis. Pelo contrário, elas são essenciais quando se deseja estabelecer um histórico, reincidência, evolução ou cronologia de um determinado aspecto ou fato. De toda forma, de pouco adianta ter informações pregressas quando o que se quer é acompanhar a efetivação do pacto, estimar prazos ou evitar percalços.

Até aqui vimos que *accountability* para IA é ser moralmente responsável pelos sistemas de IA criados, operados, mantidos etc. Isto é, a obrigação moral, profissional ou legal de responder por atos próprios, alheios, ou pela IA relacionados ao cumprimento de determinadas leis, atribuições ou funções. Portanto, oriunda de um pacto firmado entre produtores, operadores, mantenedores etc. e consumidores, investidores, apoiadores, governo, sociedade etc. a fim de provermos uma IA de confiança.

Dessa forma, *accountability* para IA envolve duas responsabilidades ou deveres principais: realizar certas ações, ou abster-se de, de acordo com as expectativas de um grupo e fornecer uma avaliação dessas ações para os interessados. Para construir estes cômputos, dadas as obrigações e expectativas, as empresas laçam mão do processo de análise factual.

A análise factual é o processo de coletar, analisar e comunicar informações para permitir que as partes interessadas tomem decisões informadas (relevante e baseada em fatos). As principais características da informação de qualidade são: relevância, confiabilidade, comparabilidade verificabilidade, compreensibilidade e disponibilidade.

A fim de prover a análise factual, precisamos de uma metodologia que viabilize o cumprimento dos diversos pactos firmados pela concedente junto aos seus diversos solicitantes. Dessa monta, no próximo capítulo analisaremos quatro questões

normativas a serem respondidas para que possamos esboçar um sistema de informação que nos auxilie na efetivação dos tratos.

3 METAMODELO DE *ACCOUNTABILITY*

Como visto anteriormente, a contabilidade pode ser entendida como o processo de identificação, medição e comunicação de informações para permitir julgamentos e decisões informadas pelos seus usuários. Os utentes, dentro e fora da empresa, precisam decidir como alocar recursos econômicos escassos. Para tentar garantir que essas decisões de alocação sejam eficientes e eficazes, os utilizadores precisam de informações econômicas e de outras naturezas. É função do sistema contábil fornecer grande parte dessas informações.

A contabilidade pode ser vista como uma parte importante do sistema total de informações da empresa. Assim, podemos ver a contabilidade como um ciclo de decisão[23] em quatro etapas, ilustrado na Figura 1 – Quatro etapas do sistema de contabilidade.

Figura 1 – Quatro etapas do sistema de contabilidade

Fonte: Albrecht, et all (2008)

Identificação das informações – identificar e capturar informações econômicas (ou telemáticas no caso da IA) relevantes;

Registro das informações – registrar as informações coletadas de maneira sistemática;

Análise das informações – analisar e interpretar as informações coletadas;

Relato das informações – relatar as informações de maneira que atenda às necessidades dos usuários.

Das quatro etapas descritas acima, dimanaram uma série de questões sobre como viabilizar o processo. Por exemplo, da primeira etapa 'Identificação das infor-

[23] Um ciclo de decisão é uma sequência de etapas usadas por uma entidade para alcançar e implementar decisões.

mações', é natural que indaguemos "Quais informações coletar?", "Por que coletar essas informações?", "Por que divulgar essas informações", "Para quem divulgar essas informações" etc.

Para apoiar o processo de identificação, registro, análise e relato das informações, vamos concentrar nossos esforços em um framework normativo de quatro questões, largamente aplicado pelas Ciências Contábeis e Administrativas (ALBRECHT, STICE, *et al.*, 2008). Esse metamodelo será, concomitantemente, correlacionado com as nossas necessidades notariais e informacionais para a IA.

A etapa de 'Identificação das informações' será abordada nos capítulos 4 e 5. As etapas de 'Registro das informações' (abordado no capítulo 6 pelos mecanismos de telemetria e pela XAU), 'Análise das informações' (tratadas no capítulo 7 pelas técnicas de XAI) e 'Relato das informações' (capítulo 8) versarão quase que exclusivamente sobre a IA, i.e., eventualmente correlacionaremos com algumas técnicas notariais.

Neste capítulo, nos dedicaremos a apresentar o Metamodelo de *Accountability*. São veiculadas as quatro questões normativas que compõem o metamodelo: "3.2.1 (QN1) Por que coletar e divulgar informações?", "3.2.2 (QN2) Quem são os stakeholders contemplados por essas informações?", "3.2.3 (QN3) Quais tipos de informações serão coletadas e quais divulgações serão feitas?" e "3.2.4 (QN4) Como as informações devem ser divulgadas?", suas respectivas justificativa e finalidade, além dos conceitos de escopo (3.3 Escopo) e recursos (3.4 Recursos). Antes, apresentaremos os sistemas de *accountability* empresarial, que alicerçam a implementação do metamodelo por um sistema de informação (ALBRECHT, STICE, *et al.*, 2008).

3.1 SISTEMAS DE *ACCOUNTABILITY* EMPRESARIAL

Como explanado anteriormente, a administração é responsável por manter evidências, incluindo documentação, para fornecer suporte razoável para suas avaliações. As informações contábeis gerenciais correntemente são produzidas para

fins de informações internas e são consideradas informações discricionárias porque não há lei que exija que sejam fornecidas à administração.

Como informações discricionárias não são necessárias, a administração deve determinar se os benefícios de receber essas informações são maiores do que os custos de produzi-las. As informações obrigatórias geralmente são produzidas com o menor custo possível para cumprir as demandas legais.

Essas evidências também permitem que terceiros considerem o trabalho realizado pela administração. Assim, a documentação é necessária para a auditoria interna corroborar as ações da administração, bem como para os examinadores externos avaliarem as afirmações sobre o controle interno e relatórios financeiros.

A documentação inclui modelos de processos de negócios, regras de negócios, manuais do usuário, manuais de treinamento, especificações de produtos, manuais de software, cronogramas, organogramas, planos estratégicos e materiais semelhantes que descrevem: operação, restrições e objetivos dos processos e sistemas de negócios. A documentação é importante por diversas razões, incluindo: treinamento, descrição dos processos e sistemas atuais, auditoria, prestação de contas, interações padronizadas e facilitação da melhoria do processo.

Uma categoria documental vital é a dos modelos de negócios (MN). Os MN fornecem ferramentas de comunicação, treinamento, análise e persuasão. São particularmente adequados para avaliar o que precisa ser mudado e planejar como fazer a mudança. Em particular, os modelos de negócios criam valor das seguintes maneiras: gerenciando a complexidade, levantando requisitos, reconciliando pontos de vista, especificando requisitos, gerenciando a conformidade, apoiando o treinamento, gerenciando e reutilizando o conhecimento.

Os Modelos de Processos de Negócios (MPN) são compostos por: Modelos de Atividade (i.e., descrevem as atividades do processo), Modelos Estruturais (i.e., as estruturas de dados) e as Regras de Negócio (que restringem e orientam as operações do processo) (ALBRECHT, STICE, *et al.*, 2008).

Os modelos de atividade (MA) descrevem a sequência do fluxo de trabalho em um processo, ou processos, de negócios. Seu propósito é representar o fluxo sequencial e a lógica de controle de um conjunto de atividades relacionadas. Os MA

são um ferramental para planejar, documentar, discutir, implementar e facilitar o uso dos MPN uma vez implementado (ALBRECHT, STICE, *et al.*, 2008).

Existe uma variedade de modelos de atividades para documentar e analisar o fluxo de trabalho nos processos de negócios, como: fluxogramas, diagramas de fluxo de dados, mapas de processos de negócios, método de modelagem funcional IDEF0[24], diagramas de atividades UML[25] e notação de modelagem de processos de negócios (BPMN)[26]. Independentemente da notação de modelagem específica empregada para descrever suas atividades, os modelos de fluxo de trabalho devem ser capazes de descrever:

(1) Eventos que iniciam, alteram ou interrompem o fluxo no processo;

(2) Atividades e tarefas dentro do processo;

(3) A sequência de fluxo entre as tarefas;

(4) Pontos de decisão que afetam o fluxo;

(5) Divisão da atividade em razão das funções organizacionais.

Os modelos estruturais (ME) descrevem as estruturas de dados e informações inerentes a um processo, ou processos, de negócios. O objetivo principal desses modelos é criar um projeto para o desenvolvimento de um sistema (e.g., banco de dados relacional) que suporta à coleta, agregação e comunicação de informações sobre os processos e facilitar seu uso após implementação.

Modelos de dados (MD) têm sido empregados para representar o conteúdo conceitual e para se comunicar com os usuários desses sistemas. Há uma variedade de notações para descrever os elementos desses sistemas, como: diagramas de Bachman[27], modelagem entidade-relacionamento[28] e diagramas de classe de lin-

---

[24] Vide (KETTINGER e GROVER, 2000).
[25] Vide (UML, 2005).
[26] Vide (BPMN, 2017).
[27] É um diagrama de estrutura de dados usado para projetar os dados com uma rede ou modelo "lógico" relacional, separando o modelo de dados da maneira como os dados são armazenados no sistema.
[28] Também chamado Modelo ER, ou simplesmente MER, é um modelo conceitual que descreve os objetos (entidades) envolvidos em um domínio de negócios, com suas características (atributos) e como elas se relacionam entre si (relacionamentos).

guagem de modelagem unificada[29]. Um modelo de dados deve ser capaz de descrever:

(1) As entidades ou coisas no domínio de interesse;

(2) As relações entre essas coisas;

(3) As cardinalidades que descrevem quantas instâncias de uma entidade podem ser relacionadas a outra;

(4) Os atributos ou características das entidades e relacionamentos.

3.1.1 Sistema de Informação Contábil (SIC)

É o sistema que registra, processa, resume, relata e comunica os resultados de transações comerciais; fornece informações (financeiras e não financeiras) e facilita a tomada de decisões. Além disso, o SIC é projetado para garantir níveis adequados de controles internos (importantes medidas de segurança para proteger a integridade de dados sensíveis) para essas transações. Visto de forma ampla, um SIC coleta, processa e relata informações consideradas úteis na tomada de decisões. O estudo dos SIC está no cerne de duas disciplinas tradicionais: sistemas de informação e contabilidade.

O SIC geralmente é a base para um sistema empresarial (SE), também chamado de sistema de planejamento de recursos empresariais (SPRE). Para a maioria das empresas, os benefícios informacionais desses dados integrados incluem maior integridade, transparência e pontualidade das informações necessárias para gerenciar com eficácia as atividades da empresa.

O SE serve como espinha dorsal para os processos de negócios internos e como uma conexão para os processos de negócios externos. Cabe ao SIC auxiliar na integração do negócio com as partes externas (e.g., fornecedores, clientes etc.).

---

[29] Vide (UML, 2005).

A interação da empresa com os fornecedores é denominada de gerenciamento da cadeia de suprimentos (GCS), e a interação com os clientes é corriqueiramente chamada de gerenciamento de relacionamento com o cliente (GRC).

Cadeia de suprimentos (ou cadeia de abastecimento) refere-se a uma rede de processos que entrega um produto acabado, ou serviço, ao cliente final. Ela abrange o fluxo de materiais, informações, pagamentos e serviços dos fornecedores de matérias-primas, passando por fábricas e armazéns, até os clientes finais dos produtos da empresa. Além disso, inclui as empresas e processos que criam e entregam produtos, informações e serviços aos clientes finais.

O software de gerenciamento de relacionamento com o cliente (CRM) é um termo que descreve o software usado para gerenciar e nutrir as interações de uma empresa com seus clientes atuais e potenciais. Frequentemente inclui o uso de ferramentas de marketing e mineração de dados para aprender mais sobre os clientes e desenvolver o relacionamento empresa-cliente. O software de CRM também inclui as seguintes funcionalidades: gerenciar vendas e marketing, atendimento ao cliente e suporte técnico após a conclusão da venda.

3.2 QUESTÕES NORMATIVAS

Voltamo-nos agora para algumas deliberações que precisam ser feitas como parte do processo de análise factual e prestação de contas do desempenho da IA. Relembrando, análise factual consiste em três etapas: coletar, analisar e comunicar informações para permitir que os usuários dessas informações tomem decisões informadas. Dessa definição, emanam quatro indagações:

- Por que a empresa decidiu coletar e divulgar informações (ou cômputos) sobre aspectos específicos da IA?
- Quem são os stakeholders contemplados por esses cômputos?
- Quais tipos de informações serão coletados e quais divulgações serão feitas para os stakeholders internos e externos, i.e., quais são suas necessidades de informacionais?

- Como essas informações devem ser divulgadas, e.g., meio e forma?

3.2.1 (QN1) Por que coletar e divulgar informações?

A primeira questão nos remete a motivação que justificará a coleta, análise e divulgação de um determinado aspecto da IA. Ao decidir quais informações coletar, vários fatores podem ser considerados: requerimentos legais, antecipação de demandas futuras, responsabilidades percebidas, aumento dos lucros, resposta à crise ou melhoria de imagem e confiança.

Requerimentos legais, em geral, são atendidos por uma questão de conformidade, cabendo a empresa apenas coletar e divulgar informações específicas. Um exemplo clássico é o direito à explicação: “[O titular dos dados] deverá ter o direito de [...] obter uma explicação sobre a decisão tomada” (EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2016).

Antecipar demandas futuras varia de simples prognosticar um anseio até prevenir a imposição de requisitos obrigatórios por lei. Frequentemente, vários grupos dentro da sociedade farão lobby junto ao governo para implementar leis que exijam que as organizações façam proclames de quesitos específicos.

Ao se antecipar, e fornecer alguma forma de *accountability* na matéria demandada, a probabilidade de que as divulgações relacionadas possam ser exigidas pelo governo será reduzida, ou limitada a organizações específicas. Um exemplo no qual nos dedicaremos futuramente é a exigência legal da geração automática de logs[30] por parte dos provedores de sistemas de IA de alto risco (COMISSÃO EUROPÉIA, 2021).

Reponsabilidades, ou impactos, percebidos compreendem a garantia, por parte da administração, que suas operações não afetem adversamente outras pessoas e, portanto, coletam, voluntariamente, uma variedade de informações sobre o desempenho da IA. Imagine, por exemplo, que você produz um reconhecedor facial para uso diverso. Melhor do que divulgar apenas métricas sobre desempenho e efi-

[30] Vide 6.3.1 Telemetria, logs.

cácia, é preferível que constem outras informações como condições de uso e aplicação, i.e., algo semelhante a bula de remédio.

Demandas específicas de stakeholders, especialmente dos grandes, não são estranhas ao meio corporativo. As organizações frequentemente coletam e divulgam informações porque certos intervenientes poderosos querem saber sobre aspectos específicos de suas operações ou da IA. A forma de manifestar essas solicitações varia de requisitos específicos a auditorias. Para ilustrar, um parceiro como o Greenpeace deve exigir métricas ambientais sobre suas IA uma vez que os treinamentos dos modelos atuais geram um consumo de energia e emissão de CO_2 cada vez maiores (KAUFMAN, 2020).

Aumento dos lucros constitui-se num fator multiface. Algumas empresas são simplesmente reacionárias e irão efetivamente "fazer a coisa certa" porque isso irá gerar lucros maiores, não porque acreditem ser conscienciosas, ou responsáveis, por ações específicas. Quer seja por essa razão, quer seja por qualquer outra apresentada até agora, o emolumento gerado pela IA quase sempre é enaltecido em qualquer proposta, especialmente as de automação, e a ausência de *Key Performance Indicator* (KPI), i.e., indicador de performance, que comprove a eficácia econômica da solução a torna natimorta.

Resposta à crise, costuma ser outra grande justificativa para coletar e divulgar cômputos sobre a IA. De vez em quando, haverá imprevistos que afetam adversamente a forma como o público percebe a organização. Há muitas evidências disponíveis de que, se a reputação ou legitimidade de uma organização é gravemente prejudicada, a organização em questão (e talvez até outras organizações dentro do mesmo setor) produzirá esclarecimentos públicos sobre a seriedade com que está conduzindo o assunto e quais iniciativas foram implementadas para ajudar a garantir que tais eventos não ocorram novamente (LANGE, LEE e DAI, 2011).

Em 2015, um desenvolvedor de software negro constrangeu o Google ao tuitar que o serviço de fotos da empresa havia rotulado fotos dele, com um amigo negro, como “gorilas”. O Google se declarou “horrorizado e genuinamente arrependido”. A empresa disse que o rótulo gorila não seria mais aplicado a grupos de imagens e que estava trabalhando em correções para o óbice (SIMONITE, 2018).

Três anos depois o Google cumpriu parcialmente sua promessa. A solução apresentada foi remover o resultado gorila, e alguns outros primatas definidos como potencialmente racistas, do classificador; passando a apresentar a mensagem “sem resultados” no lugar da classificação.

E finalmente, um dos motivos mais turrulentos: dissimulação. As organizações estão aderindo à moda do "propósito" para convencer o mundo de que estão colocando o planeta e as pessoas ante seus lucros. Se uma determinada equipe normalmente só faz divulgações quando há benefícios comerciais diretos em elaborá-los, devemos nos questionar o quanto podemos confiar em tais divulgações. Por outro lado, se o time sempre pareceu ter respeitado os direitos informacionais das pessoas sobre os vários impactos da IA, então podemos estar mais inclinados a confiar nesses proclames, particularmente os feitos voluntariamente.

À vista disso, as razões que levam uma organização a produzir cômputos sobre a IA variam de um desígnio ético em promover uma benesse socioambiental, ou pelo menos que não impacte negativamente, à um mero ganho financeiro ou salvaguarda. Mas o motivo pelo qual esta é a primeira pergunta a ser respondida é que ela impacta diretamente na próxima pergunta sobre a quem essas contas são direcionadas, como discutiremos agora.

3.2.2 (QN2) Quem são os stakeholders contemplados por essas informações?

Uma vez determinados os motivos, i.e., o porquê reportar, é natural que se estabeleça o paciente dessa ação, i.e., quem receberá as informações. Num palavreado de comunicação, o emissor é a empresa, o receptor são os stakeholders ou, aludindo a nossa ideia de pacto, a empresa é a concedente e os stakeholders são os solicitantes.

Se os relatórios produzidos visarem exclusivamente a performance finalista da IA (e.g., precisão, taxa de acerto, falso positivo etc.) e eventuais KPI de negócio, eles atenderam apenas à engenheiros, técnicos, gerência e alguns investidores. Mas e quando houver um imprevisto, uma nova cobrança social ou uma demanda jurídica?

Cabe aqui reiterar que a categoria de afetados pela IA abarca uma subclasse, os indiretos, que por vezes é negligenciada, quer por difícil identificação quer por negligência ou questões econômicas. Dessas interações imprevistas, e outras, emanam novos horizontes para a nossa definição de stakeholders. Uma abordagem ética e abrangente da relatoria, direcionaria os pareceres para os stakeholders mais afetados pela IA, independentemente da valia destes para o negócio, e se concentraria em questões como os direitos informacionais das partes.

Em última análise, para provermos *accountability* de maneira plena, precisamos que todas as partes afetadas pela IA sejam contempladas. Para atender a essas necessidades específicas que respondemos a próxima pergunta sobre os tipos de informações coletadas e divulgadas para os stakeholders externos.

3.2.3 (QN3) Quais tipos de informações serão coletadas e quais divulgações serão feitas?

O que reportar, reside em termos de que tipo de informação será divulgada e quais teses serão abordadas pelo cômputo. Para determinarmos quais dados serão coletadas precisamos antes estabelecer o que será relatado (i.e., matéria) uma vez que essas informações coletadas serão o insumo para a análise factual e, o processo de análise empregado será assinalado pelo que será relatado.

Identificar quais matérias as partes interessadas julgam responsabilidade do encarregado pela IA reportar, envolve o engajamento de ambos. Quem envolver nesse debate, dependerá das motivações para relatar, mas os stakeholders alvos da IA, gerência, engenheiros e equipe jurídica configuram uma boa partida para conversa. No entanto, se o foco for salvaguardar os grupos mais impactados pelas operações da IA, deve-se considerar a natureza das diversas resultâncias e fornecer os devidos esclarecimentos aos afetados, e quando possível, incluir soluções ou mitigações exequíveis.

Outro fator normalmente considerado é a demanda por informações específicas pelos stakeholders. Geralmente são reivindicações para verificar e/ou validar a existência ou conformidade de procedimentos ou políticas internas pelo demandan-

te, i.e., verificar se estamos fazendo o que combinamos. O exemplo clássico aqui são as análises de risco e qualidade, principalmente quando essa IA está embarcada em um veículo, e.g., carro, avião, drone.

Também, devemos considerar quando um determinado stakeholder têm necessidade por informações específicas. O impacto do que respondemos na primeira questão, aqui, é primacial, i.e., podemos apresentar análises para antecipar demandas futuras, reponsabilidades percebidas por grupos específicos ou responder especificamente a um grupo, numa eventual crise.

O que está sendo dito aqui é que, ao indagar às partes interessadas sobre quais informações desejam, às vezes se faz necessário educá-las sobre quais informações podem ser disponibilizadas e, por sua vez, como essas informações podem ser usadas para tomar decisões informadas. Não é raro, principalmente nessa área, que expectativas informacionais excedam a capacidade técnica de reportá-las ou mesmo questões legais ou estratégicas.

Um exemplo interessante envolvendo essas duas ênfases implica o algoritmo de corte de imagem do Twitter. Os usuários notaram que quando duas fotos - uma de um rosto negro e a outra de um branco - estavam na mesma postagem, o Twitter geralmente mostrava apenas o rosto branco no celular. O Twitter disse que testou o algoritmo para preconceito racial e de gênero durante o desenvolvimento e nenhum viés significativo entre etnias (ou gêneros) foi encontrado, mas iria revisar o estudo. Também afirmaram que muitas perguntas serão eliciadas e os detalhes serão compartilhados depois que as equipes internas tiverem a chance de examiná-las (BBC NEWS, 2020).

Contudo, nem sempre a organização pode divulgar todos os tipos de informações que dispõe sobre os vários aspectos da IA, caso contrário, no mínimo, os relatórios produzidos seriam extensos e ininteligíveis, ferindo os princípios qualitativos da relevância, compreensibilidade e, parcialmente, da disponibilidade. Em vez de fornecer informações sobre cada impacto diferente, a organização pode fornecer detalhes, dos efeitos mais significativos, juntamente com explicações de porque são considerados os mais importantes de monitorar e controlar. Quando se fizer necessário reportar pormenorizadamente, deve-se incluir um sumário executivo, que fará as vezes de resumo dos principais tópicos do documento.

Vale evocar aqui uma observação sobre transparência. As deliberações sobre para quem divulgar os cômputos, e o que publicizar, afetam diretamente a percepção sobre a isenção destes. As informações divulgadas serão assimiladas como fundamentalmente equilibradas (confiáveis) ou serão meramente informações predominantemente favoráveis ou positivas e, portanto, tendenciosas? E para suplantar esse e outros óbices que respondemos a próxima pergunta sobre como divulgar as informações.

3.2.4 (QN4) Como as informações devem ser divulgadas?

Para entender como as informações devem ser divulgadas, precisamos determinar se existe uma estrutura de relatório apropriada e onde as informações relacionadas devem ser divulgadas. Do contrário, ou se acharmos que as estruturas disponíveis não atendem às demandas ou necessidades das partes interessadas, podemos criar nossos próprios relatórios.

Também, precisamos considerar onde as informações devem ser divulgadas. De fato, existem muitas opções no que diz respeito à melhor mídia para disponibilizar informações às partes interessadas.

Embora os processos envolvidos na coleta e relato dos vários aspectos do desempenho da IA possam ser de natureza iminentemente técnica, a maneira como os recebedores dos diversos cômputos usa os dados pode, em última instância, gerar vários impactos sociais. Se assim não fosse, a confecção dos cômputos seria fundamentalmente nula.

Por exemplo, usuários de IA explicáveis tomarão várias decisões com base nas informações fornecidas. Imagine o seguinte cenário: você decide por sua casa a venda e descobre que a imobiliária escolhida possui um “avaliador virtual”. Segundo o site, basta informar alguns dados, e.g., endereço, área construída, número de banheiros etc., e o adjutor exibe a cotação. Se você apenas almeja por uma estimativa do seu imóvel, provavelmente você estará satisfeito. Contudo, se você não concordar com a avaliação, e estivesse na presença de um avalista humano, você o inda-

garia por uma justificativa da peritagem, i.e., quais fatores pesaram para aquele resultado.

Uma IA explicável no exemplo acima nos apresentaria a estimativa e as razões, i.e., as justificativas, que a levaram a concluir o valor apresentado, o que nos ajudaria não só a entender o valor, mas obter, por exemplo, dicas do que melhorar no imóvel. Algo como: “Casa de grande qualidade, recém reformada, em boa vizinhança, com garagem e três pisos, estas casas estão com procura em alta no mercado. Contudo, endereço não coberto pela rede de gás, piscina sem manutenção, sem equipamentos de segurança. Por causa disso, o preço previsto é...”.

Os cômputos sobre imparcialidade podem demonstrar que a IA apresenta um comportamento anômalo em determinada situação. Como vimos anteriormente, o algoritmo de corte de imagem do Twitter, apesar de ter sido testado e aprovado para preconceito racial e de gênero durante o desenvolvimento, apresentava um comportamento divergente da moral hodierna. Essa constatação levou inúmeros usuários e reclamarem e seus anseios foram ouvidos (BBC NEWS, 2020).

Similarmente, a organização pode produzir avaliação que mostra a aprovação da IA em uma bateria de testes de resiliência, o que impacta diretamente na sua percepção de qualidade e segurança. Essa é uma característica que geralmente é observada pelo consumidor de bens duráveis, e.g., carro, barco, avião, imóveis etc., e costuma ser um critério de escolha ou desempate.

Mas numa IA? você se pergunta. Suponha que você está cogitando adquirir um telefone novo, desses com reconhecimento facial e de voz, que basta um simples sorriso ou palavra para desbloquear ou autorizar pagamento. Você se sentiria entusiasmado em adquiri-lo, sabendo que com um simples par de óculos e fita adesiva, ou um gravador de bolso, um atacante pode desbloqueá-lo? (FISCHBACH, 2019) Ou que a IA que dirige seu carro autônomo não sabe a diferença entre o céu e um caminhão tombado na pista (記者林宜樟, 2020)? Também pode haver interações entre as várias medidas de desempenho.

COMPAS é uma ferramenta de gerenciamento e suporte à decisão, com base estatística, desenvolvido para avaliar a probabilidade de um réu reincidir. Foi confeccionado para ser configurável pelo operador a fim de atender as idiossincrasias lo-

cais do sistema judiciário e adequação a diferentes populações. No nível elementar, os fatores são abordados isoladamente, desconsiderando a influência que os demais resultados podem exercer sobre o fator analisado. No último nível, a interpretação é totalmente integrada.

A ProPublica descobriu que os réus negros, analisados pelo COMPAS, eram muito mais propensos do que os réus brancos a serem incorretamente considerados com maior risco de reincidência, enquanto os réus brancos eram mais propensos do que os réus negros a serem sinalizados incorretamente como de baixo risco. O mesmo estudo aponta que diversos fatores considerados pelo sistema contribuíam de forma decisiva para o resultado enviesado e que, por vezes, seu emprego não se justificava em bases científicas (PROPUBLICA, 2016).

Temos aqui um caso em que ao se aclarar o raciocínio empregado pela IA pudemos estabelecer não só as causas que contribuíram para o erro de classificação junto a um grupo étnico, como pudemos estabelecer quais premissas concorreram para tal. Embora grande parte da discussão acima se refira aos impactos sociais que podem surgir como resultado de diferentes atores usarem as informações divulgadas pela empresa, o próprio ato de registrar e relatar informações específicas também pode afetar o comportamento da organização, e isso pode ter efeitos sociais.

Tomemos como exemplo uma IA em desenvolvimento pela Statistical Analysis System (SAS) para ajudar a detectar com mais precisão lesões cancerosas. A tecnologia atua como um "oncopatologista virtual" do médico e explica quais fatores em uma imagem de ressonância magnética contribuem para o algoritmo discernir entre as áreas suspeitas como prováveis zonas cancerígenas, enquanto outras não (SAS INSTITUTE, 2021).

A IA explicável permite que a máquina avalie os dados e chegue a uma conclusão, ao mesmo tempo que apresenta ao médico, ou enfermeira, o raciocínio factual, i.e., baseado em dados, para que entenda como essa conclusão foi alcançada e, em alguns casos, chegar a uma conclusão diferente, e.g., em casos que requerem interpretação humana.

Isso pode economizar muito tempo da equipe médica, permitindo que se concentrem no trabalho interpretativo da medicina, e permiti-la amparar mais pacientes e dedicá-los mais atenção. Assim, os impactos sociais podem ocorrer antes mesmo da informação ser relatada.

Esclarecemos aqui que não nos esquecemos que *accountability* se trata de um trato entre solicitante e demandado. Lembremos que idealmente esse pacto se daria entre pessoas naturais e isso acarretaria uma série de facilidades que, quando a tratativa se dá entre empresa e stakeholders, se tornam mais obtusas.

Ademais, essas quatro questões normativas visam auxiliar a fabricante (concedente) a identificar e agrupar seus possíveis solicitantes, suas demandas informacionais e a melhor forma de atendê-las. Do ponto de vista do solicitante, este possui um rol completamente diferente de questões normativas, e.g.:

- O que quero que seja feito?
- Quais resultados eu almejo?
- Como formular meu pedido?
- Como mensurar o sucesso da tarefa?
- Seria melhor dividir a tarefa em partes menores?
- Quem é a melhor pessoa para a execução da tarefa?
- Ela está capacitada para essa tarefa?

O motivo pelo qual retomamos o mote é que para operar o pacto, toda a cadeia produtiva da IA deve ser engajada. Isso quer dizer que haverá uma torrente de promessas (i.e., as tarefas das perguntas acima) e com isso os vários elos dessa corrente produtiva alternaram entre solicitante e concedente. Assim sendo, teremos uma série de novos stakeholder internos, com necessidades informacionais específicas que serão conformadas e compartilhadas de formas diferente.

Faz-se, portanto, necessário que abordemos a noção de escopo, i.e., entender quais são as expectativas do solicitante em relação ao pacto e levantar as suas características para entregar de acordo com o esperado. Outrossim, precisamos suscitar o detalhamento de todo o trabalho necessário para entregar o que se espera dentro das expectativas do demandante.

3.3 ESCOPO

Como discutimos anteriormente, as crenças da gerência (e da própria empresa) sobre suas responsabilidades conformarão a *accountability* por eles apresentada, i.e., para quem e o quê. Por sua vez, isso também afetará a natureza dos cômputos produzidos para fins de gerenciamento interno e relatórios externos.

Se o time estiver interessado apenas em coletar informações sobre o desempenho objetivo da IA, juntamente com informações sobre os recursos, e.g., dados, e obrigações da organização, e.g., legais ou normativos, em termos estritos, então o escopo do relatório pode ser restrito aos muros da empresa. Teremos, portanto, uma visão circunscrita das responsabilidades algorítmicas. Conquanto, se lhes tocam as implicações sociais, securitivas e informacionais da IA para com os demais stakeholders, inevitavelmente ampliarão o escopo da análise para fora da empresa.

Na prática, o escopo do parecer possui duas dimensões, i.e., métricas e contexto. Podemos nos referir a miríade de informações sobre o desempenho da IA (e.g., as diversas métricas sobre imparcialidade, explicações sobre a mecânica da IA ou avaliações de resiliência), bem como ao contexto sujeito a coleta e relato (i.e., demais entidades informacionais que coadjuvam a IA, e.g., outros agentes, usuários, provedores de informação etc.).

As súmulas gerenciais, i.e., denominação genérica para os diferentes cômputos produzidos com fins gerenciais, são as informações necessárias para tomada das várias decisões relativas ao planejamento, monitoramento e controle das operações da IA. Versus os informes externos, i.e., denominação genérica para os diferentes cômputos produzidos com fins explicativos, são os relatórios para uso por pessoa estranha à organização, e.g., grupos afetados pelo emprego da IA num determinado segmento.

Observamos que os informes externos são o principal alvo de regulamentação, enquanto as súmulas gerenciais são basicamente não regulamentadas refletindo diretamente a predileção da cúpula pelo que observar. Os informes externos apresentam um caráter mais histórico, e.g., previsões iniciais ou dados históricos,

em vez de serem prospectivos ou comparativos, e.g., destacando as razões para as variações entre os resultados projetados e alcançados, como as súmulas gerenciais.

Historicamente, as súmulas gerenciais visão os aspectos priorizados pela cúpula e são produzidos com periodicidade alta, variando de segundos a dias. Os informes externos costumam ser produzidos com uma frequência bem menor, podendo variar entre semestral ou anual até uma ou nenhuma vez, e são voltados para a publicação de informações para pessoas que dependem ou são afetadas pela IA mas, que não fazem parte da empresa ou da sua gestão.

No tocante a granularidade, os informes externos tendem a fornecer dados agregados, geralmente abordando externalidades e com análises mais superficiais, embora informações limitadas a respeito do desempenho de diferentes ênfases sejam fornecidas a fim de criar um contexto. As súmulas gerenciais, no entanto, são, na maioria, desagregadas e geralmente apresentam uma série de KPI voltados para uma determinada ênfase ou análise específica, i.e., variam de informações específica sobre uma instância em particular da IA, até, processos, produtos, linhas, segmentos ou negócios.

Uma perspectiva egrégia das súmulas gerenciais é fornecer uma análise *ex-ante* sobre determinada matéria enquanto os informes externos apresentam uma síntese *ex-post*. Ilustrativamente, uma coletânea de estudos sobre a imparcialidade da IA, produzidos pela empresa para os diversos gestores, pesaria para a probidade do modelo e boa-fé da entidade enquanto uma antologia de informes externos demonstraria, parcial e publicamente, conformidade do algoritmo para os temas pactuados.

Todas essas métricas aferidas e diagnósticos auferidos originaram-se da análise dos recursos empregados e produzidos pela IA. Enfoquemos agora neles.

3.4 RECURSOS

Podemos definir recurso como algo que permita a IA realizar uma atividade que a possibilite alcançar o resultado desejado. Geralmente a empresa só relatará

os recursos que usa como parte das operações da IA sob seu controle. Os recursos são considerados controlados quando se pode negar, ou regular, o acesso de terceiros a eles. Exemplos de recursos controlados que compõe as análises incluem: bases de dados, sensores e atuadores.

No caso da IA, seu principal insumo é informação, i.e., dados. Presentes na vida da IA da sua gênese ao seu despojo, geralmente apresentam-se sob a forma de imagem, áudio, texto ou dados tabulares. Não existem limites práticos hoje em dia para forma ou volume de dados que se possa coletar, acredita-se que em 2025 alcancemos cifras de 175 zetabytes (REINSEL, GANTZ e RYDNING, 2018).

A IA também se vale de muitos outros recursos que não possui ou controla para operar, tais como: recursos naturais (e.g., energia), recursos humanos (e.g., mão-de-obra técnica, mão-de-obra gerencial, capital intelectual) ou recursos / infraestruturas sociais ou públicos (e.g., estradas, Internet). Se um determinado ativo não é mostrado como um recurso de uma organização, a implicação é que o uso ou o impacto negativo sobre este normalmente não será relatado como um sumpto ou incumbência da empresa. A prática, no entanto, torna os motivos para excetuar recursos inextricáveis, variando de segredo industrial ou contratual a impossibilidades técnicas.

Independentemente da escusa, a questão é se devemos apenas registrar e relatar informações sobre o uso e possíveis danos aos recursos controlados pela organização, ou se devemos considerar os impactos sobre os recursos não pertencentes ou controlados por ela. É cognoscível atualmente esse múnus no tocante a dados, mas, sobre recursos naturais ou humanos nem tanto. A título de exemplo, deveria o Uber aferir o número de motoristas desempregados, ou mesmo se responsabilizar, quando ela passar a empregar carros autônomos? Ou num exempli mais ubíquo e tangível, deve o produtor relatar e responder por demanda copiosa de energia para confecção da IA? Deveria este comprar, eventualmente, créditos de carbono para zerar sua pegada de carbono?

Na ausência de qualquer regulamentação que exija difusão, a decisão de revelar informações sociais ou ambientais específicas é geralmente tomada pela superintendência. Em se julgando responsáveis por garantir que recursos específicos não sejam afetados adversamente por suas operações, eles aceitam tacitamente res-

ponsabilidade com relação a esses recursos. Consequentemente, devem gerar vários cômputos que lhes permita acompanhar, controlar e melhorar o desempenho relativo aos respetivos recursos.

O fato de reconhecermos esses custos nas análises dependerá de até que ponto queremos estender nossas responsabilidades, e as eventuais consequências disso, i.e., contas e esclarecimentos prestados. Retomando o exemplo do Uber, ele pode optar por escamotear eventuais custos com indenização dos motoristas nas contas de implementação da IA autônoma ou pode optar por publicizar um plano de demissão, onde ampara seus recém desligados colaboradores com uma série de incentivos, e.g., plano de capacitação, auxílio desemprego, plano de recolocação no mercado de trabalho etc.

3.5 MAIA

O MAIA pode ser encarado como um metamodelo de negócio (especificamente, nosso 'negócio' é *accountability*). Ele cria valor ao gerenciar complexidades, levantar requisitos, reconciliar pontos de vista, especificar requisitos, gerenciar conformidade, apoiar treinamento e gerenciar e reutilizar o conhecimento.

Para sua implementação, seu modelo de processo de negócio precisa descrever as atividades do processo (MA), as estruturas de dados (ME) e as regras de negócios (RN) que restringem e orientam as operações do processo.

Como as Regras de Negócios impactam diretamente nos Modelos de Atividade e Modelos Estruturais usitados, adotaremos as 4 questões normativas, trabalhadas anteriormente nesse capítulo, e ao longo dessa obra, como fonte para as restrições e orientações aos procedimentos do processo.

Como MA, adotaremos o mesmo ciclo de decisão ilustrado na Figura 1 – Quatro etapas do sistema de contabilidade. Por se tratar de um metamodelo não cabe aqui deslindar detalhes que perpassem a camada de arquitetura, por isso, o ME será tratado de forma holística no capítulo 6.

Tabela 1 – Metamodelo de Accountability para Inteligência Artificial (MAIA)

Modelo de Atividade	❶ Identificação das informações
	❷ Registro das informações
	❸ Análise das informações
	❹ Relato das informações
Modelo Estrutural	XAU
Regras de Negócio	❶ Por que coletar e divulgar informações?
	❷ Quem são os stakeholders contemplados por essas informações?
	❸ Quais tipos de informações serão coletados e quais divulgações serão feitas?
	❹ Como as informações devem ser divulgadas?

Fonte: O Autor (2023)

Neste capítulo podemos ver a contabilidade como um sistema de coleta, processamento e comunicação de informações composto por quatro etapas: identificação das informações, registro das informações, análise das informações e relato das informações. Sabedores das exigibilidades da *accountability*, em especial fornecer avaliação de acordo com as expectativas de um grupo, nos valemos do sistema de contabilidade para prover os insumos de nossa análise factual.

O meta-arquétipo de *accountability*, que pretendemos desenvolver durante essa obra, lança mão de quatro questões normativas que evolverão os dados financeiros (ou telemáticos) à sabedoria imprescindível para a tomada de decisão informada. No próximo capítulo nos aprofundaremos nas diversas ênfases que contribuem para a coleta e divulgação desses cômputos.

4 POR QUE COLETAR E DIVULGAR INFORMAÇÕES?

Neste capítulo, nos debruçaremos sobre as principais ênfases que pesam para a resposta da nossa primeira questão normativa: por que coletar e divulgar informações? Apesar dos primeiros indícios, apresentados nos capítulos anteriores, solicitantes e concedentes terão interesses distintos, e quase sempre conflituosos, sobre o que coletar e divulgar para uma mesma ênfase.

A missão fundamental de uma empresa é dar lucro, e a da IA é executar uma tarefa, ou rol de, da forma mais eficiente possível. Podemos definir missão como a razão de ser de uma empresa, ou como a finalidade da IA, i.e., o propósito pelo qual trabalham. Por exemplo, a missão da Google é "organizar as informações do mundo todo e torná-las universalmente acessíveis e úteis" (GOOGLE, 2022).

À primeira vista, a missão parece distar das atividades cotidianas, no entanto, é ela que norteia as ações diárias. Nela estão contidas a atividade primária da empresa (i.e., o que faz), o público que deseja atingir (i.e., para quem faz), o que almeja com sua atividade (i.e., qual a sua finalidade) e como realiza suas atividades (i.e., como e onde é feito). A cada definição estratégica e a cada tomada de decisão, os gestores devem avaliar se suas decisões estão alinhadas com a missão da empresa.

Se retomarmos o exemplo da Google, é um corolário de sua missão a coleta e divulgação de informações. Contudo, mesmo que sua missão, ou a da sua IA, não sejam tão explícitos quanto a coleta e divulgação de informações, o mínimo de contexto informacional é necessário para a boa administração e realização da missão (e.g., finalidade, atividades etc.) e ocasionar a visão da empresa.

A visão de uma empresa procura responder aonde a empresa quer chegar e o que deseja ser no futuro, i.e., representa os objetivos que serão alcançados a longo prazo. Usualmente, abriga objetivos de longo prazo (i.e., como a organização se vê e onde pretende chegar), metas (i.e., o que pretende alcançar), produtos e serviços futuros e imagem de mercado que deseja alcançar (i.e., como deseja ser vista por clientes e parceiros). Por exemplo, a visão da Google é “ser o buscador de maior prestígio e o mais importante do mundo, além de ser um serviço gratuito e fácil de

usar que apresenta resultados relevantes em uma fração de segundo.” (GOOGLE, 2022)

Podemos entender a visão da empresa como um pacto de longo prazo e, dessa forma, ela deve vislumbrar tanto aspirações quanto inspirações corporativas. As aspirações correspondem aos desejos que podem ser alcançados de forma objetiva e direta (i.e., objetivos de longo prazo, meta e produtos e serviços futuros), e.g., “ser o buscador de maior prestígio e o mais importante do mundo”. As inspirações, relacionam-se com o imo, desejos e anseios dos solicitantes (i.e., imagem de mercado que deseja alcançar), e.g., “ser um serviço gratuito e fácil de usar que apresenta resultados relevantes”. A visão sempre carrega em si uma série de tarefas que para sua efetivação demandam coleta e divulgação de informações.

Há ainda uma terceira qualidade que define a personalidade da empresa e orienta a atitude de seus colaboradores, os valores. Os valores são os princípios éticos, morais e normativos que regem as ações e comportamentos da empresa enquanto seus colaboradores trabalham para alcançar a visão.

Esses princípios compreendem: postura profissional (i.e., como os colaboradores devem se portar), relações interpessoais (i.e., como os empregados se relacionam entre si), atendimento aos clientes (i.e., como a empresa se relaciona com os clientes) e relacionamento com fornecedores (i.e., como a empresa faz negócios). Recentemente, outras categorias foram incorporadas ao rol de valores corporativos e.g., responsabilidade social, respeito ao meio ambiente e respeito a diversidade de gênero.

A tríade missão, visão e valores da empresa define o propósito e a identidade da organização, i.e., a cultura organizacional. Juntamente com o planejamento estratégico empresarial, i.e., documento que estrutura todas as ações do negócio em um determinado período (AIESEC, 2021), são as principais justificativas para coletamos e divulgamos informações por parte da concedente.

A divulgação pública de informações sobre as várias facetas do desempenho da IA de uma organização é, na maioria dos países, predominantemente um exercício voluntário. No entanto, os países geralmente têm requisitos legislativos para que as informações relacionadas ao desempenho da IA, ou por hora sobre os dados co-

letados e processador por elas, sejam divulgadas ao governo, inclusive em relação a: Dados Pessoais Identificáveis (*Personal Identifiable Information*, PII), finalidade da IA, tecnologia empregada etc.

Na prática, a escolha da organização em produzir relatórios públicos, ou um site com divulgações específicas, sobre suas IA é geralmente voluntária. O que pode ser contrastado com os relatórios financeiros, que são altamente regulamentados[31], principalmente para organizações específicas, e.g., setor bancário e financeiro.

Uma razão para essa diferença tem a ver com a história. Argumenta-se que a escrita surgiu na Mesopotâmia de um sistema de contagem de fichas de argila usadas para registrar transações de mercadorias. Este sistema de cômputos foi usado a partir de 7500 A.C. em todo o Crescente Fértil e foi alterado e enriquecido para refletir novas e complicadas necessidades e desejos até os atuais balanços financeiros modernos (SCHMANDT-BESSERAT, 1992). Ou seja, a história dos relatórios financeiros é muito mais longa do que a própria IA e, portanto, teve mais tempo para ser regulamentada.

Outra possível razão, correlata, é que as preocupações da comunidade sobre o desempenho da IA são fenômeno relativamente recente. Outrossim, uma possível razão para a relativa disparidade na regulamentação da divulgação é a diferença de poder entre os stakeholders demandantes da informação. Portanto, o governo pode precisar de mais tempo para determinar quais divulgações devem ser incluídas na lei.

Dentro do sistema social mais amplo, é geralmente aceito (embora não universalmente) que os dirigentes tenham responsabilidade pelos impactos da IA de suas organizações. A falha de uma organização em relatar informações sobre certos aspectos importantes do desempenho da IA pode, do ponto de vista de várias partes interessadas, prejudicar a aceitabilidade ou legitimidade da organização. Portanto, os gerentes precisam ter informações fiáveis sobre os vários aspectos do desempenho da IA de sua organização para gerenciar adequadamente suas operações.

---

[31] Tais regulamentos obrigam a divulgação de vários itens específicos relativos ao desempenho financeiro.

Por que uma organização deve relatar externamente cômputos sobre o desempenho da IA? Pode ser porque várias partes interessadas externas à organização esperam ter essas informações, e os avalistas sentem a responsabilidade de fornecê-las.

De certa forma, a avaliação também pode fazer parte de um processo de gestão de stakeholders. Ao fornecer essas informações, os gerentes podem esperar influenciar (gerenciar) o apoio que recebem de determinados grupos de stakeholders.

A decisão de divulgar informações específicas sobre a IA geralmente ocorrem após uma atenção crítica da mídia, principalmente após eventos ou crises significativas. Por vezes nos deparamos com algum porta voz fazendo divulgações de natureza reacionária na mídia. Tais divulgações parecem ser mais sobre gestão de stakeholders do que sobre a demonstração de níveis adequados de responsabilidade. Se as organizações quiserem demonstrar altos níveis de responsabilidade, elas não esperarão até que algo dê errado, ou que surja alguma crise, antes de começar a fazer divulgações.

Algumas outras razões para divulgar informações, incluem: o desejo de mitigar riscos; possibilidade de obter uma vantagem competitiva; identificar áreas potenciais de redução de custos; meio de atração e retenção de funcionários qualificados; combater quaisquer movimentos governamentais para introduzir regulamentações sobre divulgação que possam ser onerosas; porque stakeholder influentes querem as informações; porque há uma percepção dos gestores de que as divulgações ajudarão no desempenho financeiro da organização (MADHANI, 2008).

As razões por trás da produção dos balanços que fornecem informações sobre diversos aspectos do desempenho da IA são multifacetadas. Duas áreas de consideração que podem ter um efeito significativo nessas decisões são o provável impacto financeiro, supracitados, e o potencial impacto reputacional. A seguir, abordaremos algumas das principais virtudes que afetam a geração de cômputos: ética, moral, reputação, normas, leis, políticas e educação.

4.1 ÉTICA

No intuito de prover respostas a questões éticas, face a ubíqua adoção da inteligência artificial nos últimos anos, organizações internacionais estabeleceram comitês *ad hoc* de especialistas em IA. Como parte de suas atribuições institucionais, esses comitês produziram relatórios e documentos de orientação sobre IA (e.g., código de ética) além de exercerem funções de relações públicas, ombudsman e governança em algumas instituições privadas e associações profissionais.

Entre os principais comitês destacam-se: High-Level Expert Group on Artificial Intelligence, nomeado pela Comissão Europeia; Expert Group on AI in Society (AI-GO), nomeado pela Organisation de Coopération et de Développement Économiques (OCDE); Select Committee on Artificial Intelligence, da Câmara dos Lordes do Reino Unido e o Advanced Technology External Advisory Council (ATEAC), do Google.

4.1.1 Diretrizes Éticas

Fruto do trabalho dessas comissões, os relatórios, manuais e códigos de IA ética são exemplos de normas. Ao contrário das leis – i.e., regulamentos juridicamente vinculativos, aprovados pelas legislaturas para definir conduta permitida ou proibida – as diretrizes de ética não são juridicamente vinculativas, mas de natureza persuasiva (JOBIN, IENCA e VAYENA, 2019). Essa característica normativa atribui a tais documentos uma capacidade de auxiliar na tomada de decisão em certos âmbitos, observando-se uma influência comparável a das normas legislativas (CAMPBELL e GLASS, 2001). Mas estariam esses vários grupos convergindo para o que deveria ser a IA ética? Quais devem ser os princípios éticos que determinarão o desenvolvimento da IA?

Jobin, Ienca, & Vayena conduziram uma revisão de escopo da literatura existente (incluindo princípios, diretrizes e relatórios institucionais), contendo os princípios e diretrizes para normas éticas da IA, emitidas por diversas instituições (fontes

acadêmicas e legais foram excluídas). Com base em seus critérios de inclusão / exclusão, 84 fontes ou partes delas foram incluídas na síntese final.

As categorias éticas auferidas foram inspecionadas, avaliadas quanto à consistência e compiladas, gerando treze categorias éticas. Após avaliação independente, duas categorias foram combinadas com outras, devido à proximidade semântica e temática.

Nenhum dos princípios éticos, tomado isoladamente, apareceu em todos os documentos analisados, embora exista uma confluência para o seguinte conjunto de princípios: transparência, justiça e igualdade, não maleficência, responsabilidade e privacidade. Esses princípios são referenciados em mais da metade de todas as fontes. Análises temáticas adicionais revelaram divergências semânticas e conceituais significativas na maneira como os 11 princípios éticos são interpretados, como devem ser implementados e quanto aos seus objetivos fulcrais. Abaixo analisaremos cada um deles.

> **Transparência** – a transparência é o princípio mais prevalente na literatura e compreende os esforços para aumentar a explicabilidade, a interpretabilidade ou outros atos de comunicação e divulgação. O conceito, no entanto, apresenta uma variação significativa em relação à interpretação (como uma maneira de minimizar os danos e melhorar a IA), justificativa (razões legais ou para promover a confiança, o diálogo, à participação e os princípios da democracia), domínio de aplicação e modo de realização (auditorias e a auditabilidade).

Para se obter uma maior transparência, os escritos sugerem uma maior divulgação das informações por parte daqueles que desenvolvem ou implementam a IA. Contudo, as especificações sobre o que deve ser comunicado variam bastante, e.g., finalidade da IA, código fonte, base de dados usada na confecção da IA, limitações da IA, leis, atribuições da IA e possíveis impactos causados pela IA.

A conformação dessa comunicação, no entanto dá-se por meio de explicações em termos não técnicos, auditáveis por seres humanos (por terceiros ou vias técnicas) além da criação de canais de supervisão, interação e mediação com o público, e a facilitação da denúncia.

Justiça, imparcialidade, igualdade e equidade – a justiça é expressa principalmente em termos de imparcialidade, igualdade, equidade e prevenção contra preconceitos e discriminações indesejados, além da noção de monitoramento e mitigação (significativamente menos referenciada pelo setor privado).

O conceito de justiça varia de entendimento entre as diversas fontes. Alguns documentos se concentram na justiça como respeito à inclusão e igualdade da diversidade e o setor público enfatiza particularmente o impacto da IA no mercado de trabalho e da necessidade de lidar com questões democráticas ou sociais.

Outros documentos pedem a possibilidade de apelar, contestar, reparar e remediar as decisões da IA. Também enfatizam a importância de um acesso justo aos dados e aos benefícios gerados pela IA. Alertam também para o risco de vieses presentes nos conjuntos de dados. Enfatizam a importância de adquirir e processar dados precisos, completos e diversos, especialmente os que serão usados na fase de treinamento da IA.

As estratégias para a promoção e preservação da justiça podem ser tipificadas como: soluções técnicas (padrões, certificações ou codificação normativa explícita), transparência (informar e conscientizar o público sobre os direitos e regulamentos existentes), auditoria (testar, monitorar e auditar a IA), legais (desenvolver ou fomentar o direito de apelar, recorrer, reparar ou remediar) e políticas públicas (ação e supervisão governamental, participação da sociedade civil ou de outras partes interessadas nos processos sistêmicos e maior atenção à distribuição de benefícios).

Não-maleficência – a não maleficência abarca os anseios globais por uma IA segura, i.e., que nunca cause danos previsíveis ou não intencionais. Também implica em evitar riscos específicos ou possíveis danos e sugerir estratégias de gerenciamento de riscos.

O dano é interpretado principalmente como discriminação, violação físicas ou da privacidade. As principais diretrizes de prevenção de danos focam principalmente em medidas técnicas e estratégias de governança, variando de intervenções no nível de coleta e tratamento de dados até abordagens laterais e contínuas.

As medidas técnicas abrangem da avaliação de qualidade dos dados coletados, segurança e privacidade por design, até a criação de padrões específicos para o setor (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2020).

As estratégias de governança propostas incluem cooperação ativa entre as partes interessadas, conformidade com a legislação existente (e futuras) e a necessidade de estabelecer processos e práticas de supervisão, i.e., testes, monitoramento, auditorias e avaliações por unidades internas, clientes, usuários, terceiros ou entidades governamentais, orientados para padrões de implementação e avaliação de resultados.

Responsabilização e Accountability – a responsabilização e *accountability* raramente são definidos, apesar da universalização do termo IA responsável. Todavia, os termos são frequentemente associados a responsabilidade civil (*liability*), imputabilidade, obrigações e prestação de contas (*accountability*). A maioria dos escritos analisados aponta para os principais responsáveis pelas ações e decisões tomadas pela IA como sendo: desenvolvedores, designers, fabricantes e donos da IA. Um ponto ainda não pacificado é se a IA deveria ser responsabilizada de maneira antropomórfica (e.g., com a criação de uma nova pessoa jurídica) ou apenas seres humanos devem ser responsáveis por ela.

Privacidade – A privacidade apresenta um caráter dual no campo da IA ética, como um valor (a ser defendido) e como um direito (a ser protegido). Quase sempre é relacionada à proteção e segurança de dados e raramente com liberdade ou confiança. Estratégias para o fomento da privacidade classificam-se em três: soluções técnicas (e.g., privacidade diferencial, privacidade por design, minimização de dados e controle de acesso); regulamentação (e.g., conformação legal com a GDPR e LGPD, certificação e adaptação de leis e regulamentos a fim de acomodar especificidades); e solicita mais pesquisa e conscientização.

Beneficência – a beneficência é sempre definida como um conceito ético que visa a promoção do bem; no entanto, ela nunca é melhor balizada cabendo a analogias como promoção do bem-estar humano, paz, felicidade, criação de meios socioeconômicos ou prosperidade econômica para dilucidar o conceito.

Estratégias para a promoção da beneficência incluem alinhar a IA com os valores humanos, minimizar a concentração de poder ou minimizar conflitos de interesses.

Liberdade e autonomia – O conceito de liberdade e autonomia varia de entendimento entre as diversas fontes. Alguns documentos se referem à autonomia como uma liberdade positiva, especificamente a liberdade de florescer, autodeterminação, empoderamento ou o direito a não ser vigiado, manipulado ou admoestado. Afirmam que liberdade e autonomia sejam promovidas por meio de transparência e previsibilidade, da não alienação dos cidadãos, explicabilidade, e do consentimento informado.

Confiança – a confiança é tida como essencial para a criação de uma cultura colaborativa e na conquista de outros objetivos organizacionais. O conceito varia entre IA confiável, passando por princípios de design, até como garantir a confiança do consumidor (vide 4.3 Reputação). Sugestões para a construção e manutenção da confiança incluem educação, confiabilidade, *accountability*, processos para monitorar e avaliar a integridade da IA ao longo do tempo, e ferramentas e técnicas que garantem a conformidade com normas e padrões estabelecidos.

Sustentabilidade – a sustentabilidade exige que o desenvolvimento e a implantação da IA considerem a proteção do meio ambiente, melhorando o ecossistema e a biodiversidade do planeta, contribuindo para sociedades mais justas e promovendo a paz. A IA deve ser projetada, implantada e gerenciada com cuidado para aumentar sua eficiência energética e minimizar sua pegada ecológica.

Dignidade – a dignidade permanece indefinida nas diretrizes existentes; contudo, é tida como uma prerrogativa dos seres humanos, mas não dos robôs. Ela está entrelaçada com direitos humanos, com evitar danos, aceitação forçada, classificação automática e interação humano IA desinformada (i.e., interagir com um robô sem ser informado do fato). Segundo esse preceito, a IA não deve diminuir ou destruir, mas respeitar, preservar ou até aumentar a dignidade humana.

Solidariedade – A solidariedade é entendida principalmente em relação às implicações da IA com o mercado de trabalho, e.g., criação de uma rede de seguridade social forte. Também abarca a noção de: respeito a pessoas e grupos potencialmente vulneráveis; redistribuição dos benefícios oriundos do emprego da IA na sociedade e economia; e o não individualismo.

Como aponta o escritor Yuval Noah Harari em *21 Lessons for the 21st Century* (HARARI, 2008), a principal luta da humanidade no século 21 poderá ser a relevância. A IA aumentará drasticamente a produtividade e assim como a industrialização em massa criou a classe trabalhadora, a revolução da IA criará uma classe adaptada as novas tecnologias, contudo, ambos processos levam ao empobrecimento e desalento de uma parcela da população não adaptada à nova realidade vigente (i.e., uma classe *irrelevante* e ociosa).

A solidariedade, como princípio, deve prover:

(1) O compartilhamento da prosperidade criada pela IA. Implementando mecanismos para redistribuir o aumento da produtividade para todos, e compartilhar os encargos, garantindo que a IA não aumente a desigualdade e ninguém seja alijado;
(2) Avaliar as implicações de longo prazo antes do desenvolvimento e implantação de sistemas inteligentes, para que nenhum grupo se torne irrelevante. Por exemplo, deve-se conduzir uma avaliação de riscos e danos antes de implantar a IA; desenvolver instrumentos de política internacional; proibir armas autônomas letais (LAW); regulamentar empresas globais de mídia social ou mesmo proibir, preventivamente, certas implementações até que os impactos sociais e as necessidades regulatórias sejam claros.

A proporção quase equivalente de documentos produzidos pelo setor público e privado, alvos do estudo, faculta a IA ética concernente a todos. No entanto, as soluções propostas para abordar os desafios éticos levantados, divergem significativamente. Dos 11 princípios éticos elencados anteriormente (JOBIN, IENCA e VAYENA, 2019), notamos que há um certo grau de interseção, e até mesmo confusão, entre alguns desses cânones. Outrossim, as sugestões para uma efetiva implantação desses princípios quase sempre são superficiais, em parte pela natureza multifacetada dessas diretrizes.

O conceito de semelhanças de família (*Familienähnlichkeit*), proposto por Ludwig Wittgenstein, nos diz: coisas que se pensavam ligadas por uma característica essencial comum podem, de fato, ser ligadas por uma série de semelhanças que se justapõem, onde nenhuma característica é comum a todas.

A fim de ilustração, considere uma família com características comuns, e.g., constituição física, cor da pele, cor dos olhos, andar, temperamento. Cada criança herda certas características de cada um dos pais, e os filhos podem partilhar similaridades uns com os outros. Os irmãos não se assemelham uns aos outros da mesma maneira, mas, todos guardam uma semelhança entre si.

Seria tarefa hercúlea, quando não impossível, encontrar um denominador comum para todos os problemas acerca da IA ética. No entanto, os diferentes aspectos da IA ética podem ser vistos como membros de uma família que compartilham poucas, ou nenhuma, característica em comum. Ao invés disso, compartilham uma intrincada rede de similaridades sobrepujantes.

Dessa forma, a quase totalidade desses tratados analisados propõe que a IA ética deva ser enxergada como o conjunto de proteções contra uma pluralidade de problemas distintos e medidas protetivas, relacionados entre si. Veremos a seguir como esses preceitos éticos catalogados vão impactar na moral, reputação, normas e leis.

4.2 MORAL

A moralidade é um conceito complexo, sendo objeto de estudo da filosofia e teologia há séculos. Embora a palavra moral se refira a um valor individual, moral associado a bom e imoral a mau, tomaremos o termo em seu sentido mais amplo, i.e., como o conjunto de diretrizes inatas ou culturais que norteiam a tomada de decisão das pessoas ao analisarem questões de equilíbrio.

A moral afeta os dilemas sociais de várias maneiras. Ela pode nos afetar na tomada de decisão (nos trazendo ou não conforto ao cooperarmos), depois de termos tomado a decisão e durante falta (e.g., compaixão) ou após a defecção (através

de sentimentos como culpa, vergonha, orgulho). Ela é a única pressão social que leva as pessoas a “quererem” se comportar em benefício do grupo.

Um dilema social é a escolha que cada ator faz entre os interesses do grupo e seus próprios interesses. Pode ser entendida como a escolha que fazemos quando decidimos ou não, acatar as normas do grupo.

Pressão social é o meio pelo qual a sociedade garante que o indivíduo siga as normas do grupo, em detrimento de outra. O termo é utilizado para abranger tudo o que a sociedade faz para proteger a si mesma: tanto de outros membros da sociedade, quanto de não membros que interagem com ela. É o meio que a sociedade usa para reforçar a confiança intergrupo. Poderíamos dizer então que a moral é a pressão social que trabalha "quando ninguém está olhando".

As tendências evolutivas humanas rumo à confiança e cooperação se refletem em nossos códigos morais, éticos e religiosos. Estes códigos variam muito quanto à forma, da bíblia aos provérbios; mas, todos enfatizam comportamentos pró-sociais como o altruísmo, justiça, cooperação e confiança, e.g., “Faça aos outros aquilo que você gostaria que fizessem pra você”. Tradicionalmente, a religião era o contexto no qual a sociedade codificava e transmitia suas regras morais. Hodiernamente, o processo é legalizado e tem na educação sua principal forma de transmissão.

Nossas decisões morais são contextuais. Mesmo algo tão básico como “Não matarás” não é realmente de todo axiomático, i.e., livre de contexto. Quando modificamos o contexto incluindo questões como autodefesa, aborto ou eutanásia a questão ganha outra dimensão levando, por vezes, a aceitação de decisões contrárias.

Comportamentos pró-sociais como altruísmo e justiça podem ser universais; mas, sua expressão nas diversas culturas se dá de formas e em momentos diferentes. Isso não significa que a moral se desenvolve naturalmente, mas nos é ensinada. Esses modos de expressão são os chamados *memes*[32] e evoluem, sujeitos às regras de seleção natural, mas não são geneticamente determinados. Boa parte da nossa moral é fruto da nossa cultura. Por exemplo, enquanto justiça é uma caracte-

[32] *Meme* é uma unidade de evolução cultural que pode de alguma forma se autopropagar.

rística humana universal, noções de justiça diferem entre os diversos grupos humanos, com base em variáveis (tamanho e participação da comunidade religiosa).

Infelizmente, também somos uma espécie capaz de profundos atos de imoralidade. Ao longo da história, regimes totalitários têm tentado impor códigos morais a seus cidadãos, suprimindo alguns comportamentos até então aceitáveis e engendrando novas obrigações. A pressão moral falha por várias razões específicas: o comportamento individual das pessoas varia, o que pode incorrer em falta; os valores morais, por vezes conflitam com outras normas; a moral pode ser manipulada; a moral não possui escalabilidade.

A essa altura você deve estar se perguntando o que moral tem a ver com a nossa primeira questão normativa. Como exposto, a cultura organizacional será o conjunto de diretrizes que nortearão a tomada de decisão das empresas ao analisarem questões pactuadas. No entanto, cada solicitante possui um grupo distinto de critérios que orienta sua deliberação sobre o que deve ser informado. Harmonizar essas ontologias nem sempre é fácil, mas essencial para bom encaminhamento e cumprimento do acordo.

Como sabemos, as escolhas da empresa são consumadas pelos seus colaboradores e sobre estes recaem pressões corporativas e sociais. Contudo, esses cooperários nem sempre estarão alinhados com os interesses da companhia e podem denegar as preferências do grupo que representam em favor de outro (e.g., agindo como denunciantes ou sabotadores).

Para exemplificar, temos o caso de James Williams, ex-funcionário do Google, que decidiu denunciar como atenção conquistada pelos algoritmos da empresa colocam em risco os propósitos morais dos usuários (BARROS, 2021). Ao constatar que soluções significativas para o problema da atenção retida pelas redes sociais eram preteridas em função da maximização das receitas obtidas com visualização de propaganda, James decidiu que deveria expor as práticas da empregadora.

4.2.1 IA Moral

Até aqui nós cuidamos das interações humanas e do papel da moral na sua mediação. Entretanto, à medida que o mundo se torna cada vez mais tecnologicamente avançado e globalizado, as consequências das limitações morais humanas se tornam mais profundas. Agentes autônomos estão começando a interagir com os seres humanos regularmente. Esses agentes autônomos prometem muitos serviços que beneficiarão a sociedade, mas também levantam preocupações significativas[33].

Desenvolvimentos recentes nas áreas de computação pervasiva e inteligência ambiental, nos permite cogitar a criação de uma inteligência artificial moral. Ela seria desenvolvida para direcionar agentes autônomos ou, para ajudar agentes a superar suas limitações naturais.

Para desenvolver uma IA moral para direcionar agentes autônomos, precisamos levar em consideração que nem todos os membros da sociedade estão dispostos a segui-la. Assim, ela deve lidar com novas situações que não foram antecipadas ou ensinadas no seu desenvolvimento e lidar com uma variedade de respostas diferentes que as pessoas apresentam em situações morais (SAVULESCU e MASLEN, 2015).

Uma vez capaz de direcionar agentes, o próximo passo seria ajudá-los a superar suas limitações naturais. Para tal a IA moral, assim como nós, deve conter e operar sobre o conceito de contexto. Dessa forma ela monitoraria constantemente as condições fisiológicas, mentais e ambientais do agente. Ao fazer isso temos a capacidade de alertar o agente sobre fatores que podem impactar em seu comportamento e na qualidade de sua tomada de decisão. O passo seguinte seria auxiliar o agente em determinar e escalonar suas metas morais.

Dotada dessas três capacidades, podemos desenvolver as funcionalidades de análise e aconselhamento moral. Na função de análise moral, o agente apresenta um dilema e a IA moral iniciaria um processo de análise, i.e., questões relevantes, para ajudar o agente na deliberação moral (e.g., o problema do bonde). Já a função

---

[33] Vide ilustrações 1 e 2

de aconselhamento moral, permitiria ao agente pedir à AI conselhos morais sobre o curso de ação que ele deve tomar.

O problema do bonde, é um paradoxo moral apresentado pela primeira vez por Philippa Foot e posteriormente expandido por Judith Jarvis Thomson. Nele um bonde está fora de controle em uma estrada. Em seu caminho, cinco pessoas amarradas na pista por um filósofo malvado. Felizmente, é possível apertar um botão que encaminhará o bonde para um percurso diferente; mas ali, por desgraça, se encontra outra pessoa também atada. Deveria apertar-se o botão? (STURGEON, HARMAN e THOMSON, 1998).

O experimento *Moral Machine* adaptou o problema original em nove cenários diferentes e os resultados mostraram que as respostas das pessoas estão altamente correlacionadas com aspectos culturais e econômicos. Os pesquisadores acreditam que os dados não apenas forneceram uma visão das prioridades éticas de diferentes culturas, mas ajudarão a traçar o perímetro ético da IA em breve.

Moldamos, até aqui, a IA moral para promover e auxiliar o agente à moralidade. Uma última funcionalidade desejada seria a proteção contra a imoralidade alheia. Essa funcionalidade é de longe a mais utópica, mas essencial na operacionalização de outra forma de pressão social, a reputação.

Essa especificação de IA moral que acabamos de descrever poderia ser aplicada no auxílio e coordenação de agentes inteligentes, empresas e seres humanos. Uma IA moral auxiliaria pessoas a alcançar suas metas morais (de ações de caridade a preservação do meio ambiente), empresas com suas metas de conformidade e agentes inteligentes em decisões morais (e.g., carro autônomo no experimento da *Moral Machine*). Fato é que a relação de coexistência entre humanos e IA ainda está no começo, e seu uso para regular ações autônomas e humanas nos causa espécie.

Os seres humanos são uma espécie social, e frequentemente somos observados. Na medida em que o grupo aumenta, i.e., ganha escala, e se torna mais anônimo e heterogêneo, diminui o grau de empatia entre seus membros. Isso faz uma enorme diferença porque viabiliza outra forma de pressão social, dessa vez exercida principalmente pelos pares, a reputação.

4.3 REPUTAÇÃO

Sabemos que os indivíduos, empresas e IA com quem interagimos possuem algum tipo de reputação. Enquanto a moral é parte da razão pela qual cooperamos uns com os outros, temos várias hipóteses que apontam a reputação como principal motivo para cooperarmos, tais como o desejo de recompensa (engajamento) e o medo de punição (exclusão) por outros membros do nosso grupo.

Levamos a nossa reputação muito a sério, e gastamos muito tempo e esforço para mantê-la e defendê-la. Por esse motivo, conservamos nossa reputação ilibada, encobrimos as máculas, ou, as fraudamos completamente, em razão de um bom nome. Lembramo-nos de informações negativas sobre as pessoas, de forma mais vívida, com mais detalhes, e por mais tempo, do que de informações positivas. Do ponto de vista evolutivo, saber quem vai faltar conosco é mais importante do que saber em quem confiar[34].

Somos igualmente bons em arrolar a reputação alheia. Estudos (WILSON, 1914; BURT e KNEZ, 1995) sugerem que a fofoca surgiu como um mecanismo para descobrir sobre a reputação dos outros e, consequentemente, em quem confiar. Ela funciona como um sistema de pressão social que ajuda as pessoas a se manterem na linha.

A reputação é um mecanismo que aumenta os benefícios de cooperar e os custos de defecção. As principais forças que operam nesse sistema são: O arrependimento e o perdão. Se você errar por acidente, peça desculpas, faça as pazes, e volte a cooperar. E se alguém falta com você, perdoe e volte a cooperar.

O problema com a reputação é que ela não escala muito bem. Podemos reconhecer 1.500 rostos, mas o número de pessoas cuja reputação conhecemos é muito inferior, talvez 500 ou mesmo 150.

Uma das maneiras de escalar a reputação é a generalização, baseada em grupos específicos, i.e., clubes, associações, agremiações e consociações. A confiança é inversamente proporcional ao aumento da diversidade étnica em um grupo.

---

[34] Vide (NOWAK e SIGMUND, 2005)

Não é surpresa, portanto, a existência de uma grande variedade de atributos de consociação usados para determinar quem é como nós, e.g., cor da pele, descendência, renda etc. Marcadores de consociação são mais difíceis de assimilar e fraudar em grupos de cooperação duradoura, e.g., dialetos, que em grupos de cooperação breve, e.g., regras de etiqueta.

No entanto, aferir a reputação por meio de agremiações geralmente nos leva a casos de preconceito. Também não é suficiente querer cooperar. No entanto, você precisa saber como cooperar de acordo com as normas sociais vigentes no grupo, para que possam saber, i.e., medir, que você é cooperativo e confiável. Essas normas sociais nos dizem como cooperar com e para o grupo. Por esse motivo, as sociedades tendem a homogeneizar sua moral.

Comprometimento – Podemos compensar a falta de reputação ao nos comprometermos com uma ação, de uma maneira que não podemos voltar atrás. Um mecanismo similar é eliminar suas alternativas; dessa forma, você não tem escolha, se não se manter fiel ao compromisso. Uma segunda maneira de demonstrar comprometimento é construí-lo gradativamente. Esse passo-a-passo ajuda ambas as partes a confiarem entre si, durante o processo, uma vez que o custo de falhar é parcelado. Uma terceira maneira é através de rituais. Rituais funcionam por dois motivos: O primeiro é que sua reputação fica em jogo, e o segundo é que a sociedade provê uma série de sanções para quem os renega. Claro que seu efeito se restringe a quem os conhece e compreende.

Branding – Similar ao comprometimento, *branding* não é necessariamente sobre qualidade, é sobre semelhança. É encontrar uma conexão única, capaz de fazer as pessoas intuitivamente, visceralmente, psicologicamente se relacionarem com a sua marca numa perspectiva mais pessoal e menos lógica. A marca, então, é uma fusão metafórica entre as histórias que as pessoas têm de um produto e suas próprias histórias. A identidade da marca é a proposição de valor que a empresa faz para seus clientes, i.e., sua promessa. A publicidade pode ter por finalidade convencer os consumidores a associar uma determinada marca a certa reputação, e.g., o predomínio de famílias felizes em comerciais de margarina ou a presença de médicos e/ou especialistas em

produtos de higiene. Isso significa agregar a reputação individual à reputação de um grupo, o que beneficia todos os membros da coalizão.

Sistematização – Sistematizar a reputação nos permite confiar em uma entidade, i.e., sistema, em vez de ter que confiar em uma pessoa ou empresa. O Número de Dunbar[35] nos diz que existe um limite para o número de pessoas que podemos conhecer suficientemente bem, para decidir se devemos ou não confiar; a decisão de confiar num sistema pode servir como um fiduciário para milhões de decisões individuais de confiança.

Este é um enorme desenvolvimento nos mecanismos de pressão social, que permitiu à sociedade escalar globalmente. Por outro lado, enquanto estes sistemas de reputação têm sido um enorme sucesso, trouxeram consigo um novo tipo de quebra de confiança.

Reputação não é um sistema de pressão social eficaz, a menos que tenha consequências, i.e., recompensamos os que cooperam e punimos os que não. Recompensamos os cooperadores o tempo todo através de nossas ações. A característica comum em toda essa recompensa é a participação. Se a participação é a recompensa canônica, a exclusão é a punição correspondente. Outras punições são menos graves: violência física, danos à propriedade, e assim por diante. Por vezes, essas formas alternativas de punição são denominadas vingança. Outra punição reputacional importante é a vergonha. As punições informais são tão arraigadas em nossa sociedade que, por vezes, nos passam desapercebidas sendo certamente mais prevalentes na infância, e.g., ostracismo social na escola. No tocante ao efeito punitivo, vale relembrar mais uma vez seu caráter contextual, uma vez que o aumento do número de indivíduos não cooperativos no grupo desvirtua as práticas punitivas. Uma última ideia que vale ser mencionada é a de transferência de culpa e a figura do bode expiatório[36]. Prática que costuma ser empregada por empresas, especialmente num primeiro momento, que culpam seus funcionários diretamente responsáveis no intuito de se eximirem das suas responsabilidades (e.g., a culpa é quase sempre atribuída, primeiramente, ao piloto num acidente aéreo).

---

[35] Define o limite cognitivo teórico do número de pessoas com as quais um indivíduo pode manter relações sociais estáveis (DUNBAR, 2010).

[36] A expressão tem a sua origem no ritual proferido por Aarão que, ao pôr as mãos sobre a cabeça de um bode, transmitiu para o animal todos os pecados do povo de Israel (PORTO EDITORA, 2022).

A moral e a reputação são as duas formas de pressão social mais primitivas de que dispomos. Mesmo assim, a reputação falha em inúmeras circunstâncias.

Ocultação – Detratores tomam medidas para ocultar fatos que podem prejudicar suas reputações, ou manipulam os fatos para melhorar suas reputações. Golpistas gastam um bom tempo construindo suas reputações junto às suas vítimas. Eles empregam todo o tipo de artimanhas, fachadas e até outros partícipes para convencerem suas vítimas de que eles têm uma boa reputação, se mostrando dignos e suscitando confiança.

Minimização – As pessoas tentam minimizar os efeitos de sua má reputação. Corporações trabalham para minimizar os efeitos da má reputação sobre suas marcas, através da publicidade e relações públicas.

Indiferença – Algumas pessoas, simplesmente, não se preocupam com sua reputação. Assim como a moral, o interesse individual sobre a reputação varia. Isso, claro, é contextual. Todos nós temos diferentes reputações, por diferentes motivos, nos diferentes grupos a que pertencemos.

Erro – Algumas pessoas acabam associadas a uma reputação errada. Mesmo se alguém não faz nada errado, não há garantia de que sua reputação é merecida.

Subgrupos – Desertores se agregam em subgrupos que têm diferentes regras de reputação. Os membros de uma súcia, por exemplo, têm uma reputação terrível junto à comunidade em geral; mas, eles se preocupam, principalmente, com a sua reputação dentro de suas gangues.

Custo de oportunidade – O valor da deserção pode valer o dano à reputação. Isso é muito frequente em situações de transação única, em que as partes transigem uma única vez e não voltam a interagir, e.g., compra de ingresso junto a cambista.

Na prática, a dinâmica reputacional, especialmente quando envolve a IA, costuma envolver um comprometimento gradativo com emprego massivo de técnicas de branding e sistematização, quando cabíveis. Quando há falha, costuma ser mitigada pelas dinâmicas da indiferença, ocultação, minimização e transferência de culpa.

Tay (acrônimo de thinking about you) foi um *chatbot* desenvolvido pela Microsoft em 2016 que causou uma grande celeuma ao postar uma série de comentários ofensivos no Twitter, o que acarretou seu encerramento no dia seguinte ao seu lançamento. Originalmente desenvolvida para melhorar o atendimento ao cliente em seu software de reconhecimento de voz, Tay se comunicava através do twitter e suas respostas eram modeladas a partir do que ela aprendia com as pessoas com as quais interagia.

Os desenvolvedores da Microsoft ainda tentaram moderar as mais de 96.000 postagens; mas, após sofrer um "ataque coordenado por um subconjunto de pessoas" que "explorou uma vulnerabilidade " (LEE, PETER, 2016), Tay foi encerrada com apenas 16 horas de vida. Apesar de ter emitido uma nota onde lamenta o ocorrido e ter explicado as causas que levaram Tay a surtar, tanto a empresa quanto a tecnologia de *chatbots* sofreram abalos em suas respectivas reputações.

É interessante notar, no entanto, que enquanto tecnologias como a inteligência artificial do Bradesco (BIA), e os demais robôs com quem nos deparamos nos *call centers* da vida, expiam da mesma pecha e reputação, os assistentes pessoais como a Siri e a Alex, que já tiveram seus percalços, são vistos como epítomes da tecnologia e alta conta. Outro ponto é quanto ao ganho que poderíamos ter, em situações como essas, ao dispormos de IA moral.

O principal fator para a derrocada da reputação é o crescimento do grupo. Paradoxalmente, assistido pela tecnologia (especialmente redes sociais), a pressão da reputação pode ganhar o mundo de forma célere. No entanto, esse sistema pode padecer de imprecisão, vide as atuais guerras de narrativa e fake News, nem sempre correspondendo à realidade.

A reputação, como pressão social, funciona melhor dentro de um grupo de pessoas que se conhecem. Uma vez que o tamanho do grupo cresce e, os laços sociais entre os membros enfraquecem, a reputação por si só não basta.

4.4 NORMAS

Normas, de forma geral, são parte de uma categoria maior de regras que prescrevem, demandam, ou exigem que certas ações sejam realizadas. Elas são utilizadas de forma descritiva para se referir a comportamentos, hábitos e práticas, sem qualquer expectativa subjacente ou sanção.

Normas Injuntivas se referem ao que as pessoas acreditam que deve ser feito, o que é socialmente aprovado ou reprovado. Não só a maioria das pessoas segue, mas o faz porque acredita que é um dever.

Os quatro elementos da norma (CALABRICH, 2008):

(1) um elemento prescritivo "dever";

(2) um sujeito - aquele sobre o qual a obrigação expressa na norma recai;

(3) um ato - a ação prescrita na norma; e

(4) uma condição de aplicação - as circunstâncias nas quais o ato prescrito na norma se aplica ao sujeito da norma.

Por exemplo, “a proibição de trabalho noturno, perigoso ou insalubre a menores de dezoito anos e de qualquer trabalho a menores de 16 anos, salvo na condição de aprendiz, a partir de quatorze anos.”[37]

Algumas normas se originam de regulamentação por autoridades (ou instituições), enquanto outras parecem ter evoluído junto a culturas, histórias e tradições políticas, e, na verdade, são partes constitutivas delas. Normas que orientam nossas vidas são, em sua maioria, arraigadas em sistemas sociais, morais, jurídicos e políticos.

Os modelos existentes para normatização da IA se enquadram em uma ou mais das seguintes categorias:

Legal – o design, desenvolvimento e implantação da IA devem aderir as leis e regulamentos aplicáveis;

---

[37] BRASIL, Constituição Federal de 1988, art. 7º, XXXIII.

Ético – a IA deve buscar beneficiar toda a humanidade e não deve resultar em danos desproporcionais. Valores como diversidade, imparcialidade e benefício social devem ser inseridos de forma ética;

Moral – deve haver mecanismos integrados à IA que viabilizem todas as opções acima. Garantir a *accountability* inclui medidas técnicas, como explicabilidade, segurança e salvaguardas, integradas ao desenvolvimento de algoritmos e que auxiliam engenheiros e programadores na compreensão do comportamento da IA;

4.4.1 Princípios Normativos

Numa abordagem diferente da que vimos anteriormente no item 4.1 Ética, ao estender os princípios da ética biomédica, (FLORIDI, COWLS, *et al.*, 2018) nos apresenta as principais oportunidades e riscos que as tecnologias de IA trazem à sociedade. O modelo normativo ético proposto é composto por cinco princípios éticos - que devem fundamentar o desenvolvimento e a adoção da IA - e 20 recomendações concretas - para avaliar, desenvolver, incentivar e apoiar a boa IA.

Para embasar os princípios e recomendações sugeridos, (FLORIDI, COWLS, *et al.*, 2018) nos apresentam 4 perguntas: “quem podemos nos tornar?”; “o que podemos fazer?”; “o que podemos alcançar?”; e “como podemos interagir?”. Para responder a cada uma dessas perguntas, eles sugerem quatro oportunidades e seus desafios correspondentes.

Para a primeira, quem podemos nos tornar, eles sugerem propiciar a autorrealização humana, sem desvalorizar as habilidades humanas. A segunda, o que podemos fazer, pode ser alcançada através do aprimoramento da agência humana, sem remover sua responsabilidade.

A terceira, o que podemos alcançar, dá-se por aumentar as capacidades sociais, sem redução da intervenção humana. E a última, como podemos interagir, faculta-se pelo fomento da coesão social, sem erodir a autodeterminação humana.

Juntas, essas quatro oportunidades e seus respectivos desafios apresentam um panorama do impacto da IA na sociedade e nas pessoas. Com base no panorama obtido, 6 fontes foram analisadas, resultando em 47 princípios (FLORIDI, COWLS, *et al.*, 2018).

Os princípios éticos auferidos foram inspecionados e avaliados quanto à consistência. As fontes apresentaram muita coerência e sobreposição entre os princípios. O conjunto foi então comparado com os quatro princípios básicos comumente usados em bioética: beneficência, não maleficência, autonomia e justiça. Da análise comparativa surgiu a necessidade de mais um princípio, ausente na bioética, que foi nominado explicabilidade (i.e., inteligibilidade e *accountability*).

> **Beneficência**: Promover o bem-estar, preservar a dignidade e sustentar o planeta. Beneficência pode ser entendida como “faça apenas o bem”. O conceito enfatiza a importância central de promover o bem-estar das pessoas e do planeta.
>
> **Não-maleficência**: Privacidade, Segurança e “Capability Caution”. Não-maleficência pode ser entendida como “não cause o mal”. O objetivo é simplesmente impedir que surjam danos, seja pela intenção dos seres humanos ou pelo comportamento imprevisível das máquinas, i.e., Capability Caution, (incluindo estímulos não intencionais do comportamento humano de maneiras indesejáveis).
>
> **Autonomia**: O poder de decidir (se deve ou não decidir). Autonomia pode ser entendida como a faculdade dos indivíduos terem o direito de tomar decisões, por si mesmos, sobre o tratamento que recebem ou não. No contexto da IA o princípio da autonomia significa encontrar um equilíbrio entre o poder de decisão que retemos para nós e o que delegamos a agentes artificiais.
>
> **Justiça**: Promover a Prosperidade e Preservar a Solidariedade. Justiça pode ser entendida como a melhor (i.e., observando-se princípios de imparcialidade, igualdade, equidade, proporcionalidade, inclusão e indiscriminação) distribuição de recursos num dado contexto.

No contexto da IA o princípio da justiça significa usar a IA para corrigir erros do passado, i.e., eliminar a discriminação injusta, garantir que o uso da IA crie bene-

fícios compartilhados, ou pelo menos compartilháveis, e prevenir a criação de novos danos, i.e., o enfraquecimento das estruturas sociais existentes.

> **Explicabilidade**: Prover os outros princípios por meio da inteligibilidade e *accountability*. Explicabilidade pode ser entendida como a aplicação dos conceitos de inteligibilidade, i.e., "como funciona?", e o sentido ético de *accountability*, i.e., "quem é responsável pela maneira como funciona?" na esfera dos processos de tomada de decisão da IA. É interessante notar que a bioética relega a outras disciplinas, e.g., biologia, ecologia, fisiologia..., o encargo de prover a explicabilidade necessária.

Os cinco princípios apresentados acima congregam a acepção de cada um dos 47 princípios originalmente arrolados, formando uma estrutura ética dentro da qual serão apresentadas as 20 recomendações abaixo. Essas recomendações serão agrupadas em quatro categorias normativas: avaliação, desenvolvimento, incentivo e suporte (FLORIDI, COWLS, *et al.*, 2018).

> **Avaliação**:
>
> (a1) Avaliar a capacidade das instituições existentes, como o judiciário, de corrigir erros ou danos causados pelos sistemas de IA;
>
> (a2) avaliar quais tarefas e funcionalidades de tomada de decisão não devem ser delegadas à IA;
>
> (a3) avaliar a aderência das regulamentações atuais a ética e a capacidade da estrutura legislativa em acompanhar o ritmo dos desenvolvimentos tecnológicos.
>
> **Desenvolvimento**:
>
> (d1) desenvolver uma estrutura para aprimorar a explicabilidade dos sistemas de IA que tomam decisões socialmente significativas;
>
> (d2) desenvolver procedimentos legais apropriados, que melhorem a infraestrutura de TI do sistema judicial, para permitir o exame minucioso de decisões algorítmicas em tribunal;

(d3) desenvolver mecanismos de auditoria (que identifiquem consequências indesejadas) para sistemas de IA e mecanismo de solidariedade (que lidem com riscos graves) para setores altamente impactados pela IA;

(d4) desenvolver processo ou mecanismo de reparação para remediar ou compensar erros ou queixas causados pela IA;

(d5) desenvolver métricas comuns para mensurar a confiabilidade dos produtos e serviços de IA, para serem utilizados pelas diversas entidades de produção, manutenção e controle de IA;

(d6) desenvolver uma agência de supervisão, responsável pela proteção do bem-estar público, para avaliação e supervisão científica de produtos, programas, sistemas ou serviços de IA;

(d7) desenvolver um observatório de IA;

(d8) desenvolver instrumentos legais e modelos contratuais para estabelecer as bases para uma colaboração homem-máquina satisfatória no ambiente de trabalho.

Incentivo:

(i1) incentivar financeiramente o desenvolvimento e o uso de tecnologias de IA socialmente adequadas (i.e., aderente aos princípios éticos aduzidos) e ambientalmente amigáveis;

(i2) incentivar financeiramente pesquisas científicas no campo da IA que sejam sustentáveis, coerentes e ampliadas;

(i3) incentivar a cooperação interdisciplinar e intersetorial e o debate financeiro sobre as interseções entre tecnologia, questões sociais, estudos jurídicos e ética;

(i4) incentivar financeiramente a inclusão de considerações éticas, legais e sociais em projetos de pesquisa em IA;

(i5) incentivar financeiramente o desenvolvimento e o uso de zonas especiais, legalmente desreguladas, para o teste empírico e o desenvolvimento de sistemas de IA;

(i6) incentivar financeiramente a pesquisa sobre a percepção e compreensão do público sobre a IA e suas aplicações e a implementação de mecanismos estruturados de consulta pública para projetar políticas e regras relacionadas à IA;

Suporte:

(s1) apoiar o desenvolvimento de códigos de conduta autorregulatórios para dados e profissões relacionadas à IA, com deveres éticos específicos;

(s2) apoiar os conselhos administrativos corporativos a assumirem a responsabilidade pelas implicações éticas das suas IA;

(s3) apoiar a criação de currículos educacionais e atividades de conscientização pública sobre o impacto social, legal e ético da Inteligência Artificial.

Esse conjunto de recomendações deve ser visto como um documento em evolução. As normas elencadas são projetadas para serem dinâmicas, exigindo não apenas políticas únicas ou investimentos pontuais, mas sim esforços contínuos para que seus efeitos sejam sustentados, i.e., incorporados à cultura da empresa e dos funcionários. Caso contrário, os esforços das empresas em estabelecer normas e conselhos éticos podem ser vistos não como tentativas legítimas e efetivas de autorregulação da IA e sim como meras estratégias de relações públicas, tentativa de amenizar questões sociais como problemas técnicos, i.e., "lavagem ética", ou para evitar completamente a regulação.

A autorregulação, tão propagada pelas gigantes da tecnologia, é muito mais facilmente dita que implementada. O episódio com o Conselho de ética para IA do Google nos apresenta algumas lições pertinentes.

O Advanced Technology External Advisory Council (ATEAC), fundado em março de 2019 pelo Google, recebeu a tarefa de criar e implantar a prática de desenvolvimento responsável de IA na cultura da empresa. Após a renúncia de um membro do conselho no Twitter e uma petição, assinada por vários funcionários e acadêmicos, pedindo a remoção do presidente, o ATEAC foi dissolvido (WALKER, 2019).

Atualmente, gigantes como Google e Facebook estão cada vez mais enfrentando reação por parte de seus próprios funcionários (BERNARDINO, 2019). Além disso, a confiança dos usuários nessas plataformas se deteriorou em consequência dos sucessivos vazamentos e violações de dados (O GLOBO, 2019).

Fica evidente portanto que a mera definição de um código de ética, uma normativa associada e um eventual conselho, não bastam para a efetiva prática dessas premissas. Faz-se necessária a internalização e operacionalização desse código pelos vários extratos sociais presentes na empresa além da aderência e conformidade da normativa ética com os valores presentes na sociedade em que está inserida. Talvez caiba a necessidade de uma pressão social institucionalizada, uma instituição que assegure a devida implantação, execução e manutenção desses princípios.

4.5 LEIS

A história poderia ser vista como um processo de formalização cada vez maior das regras informais e dos mecanismos de governança necessários ante a crescente complexidade da especialização e divisão do trabalho (JÜTTING, 2003). Leis, regulamentos e normas em geral são todas pressões sociais institucionalizadas semelhantes a reputação, só que formalizadas. Talvez seja o medo da punição que nos torna livres das tentações, “mas quando os bons costumes estão faltando, a lei torna-se imediatamente necessária” (MACHIAVELLI, 1531).

Não é sabido, exatamente, quando na história os costumes sociais informais se tornam regras, i.e., uma clara distinção entre as normas explicitamente codificadas, estabelecidas pelo estado, e as informais, acordadas pelo grupo. Ainda assim, a codificação do sistema reputacional em leis, foi um grande passo para o desenvolvimento social, viabilizando agrupamentos sociais maiores e mais complexos, e.g., cidades, metrópoles e megalópoles[38].

[38] Vide (GERMANN, DAY e GALLATI, 1985)

Essas leis exigem uma instituição que assegure sua devida implantação, execução e manutenção. Essas instituições implementam regras, leis, decretos e sanções para quem as desobedece e os incentivos para quem as obedece. Uma lei bem escrita combinada com a aplicação adequada, aumenta os custos para o infrator, ao ponto onde ele é forçado a arcar com os custos integrais de suas ações[39].

Alguns dilemas sociais são particularmente resistentes à pressão institucional. A pirataria, o sequestro e a extorsão são exemplos disso. É ruim para a sociedade pagar resgate – pela carga, pessoa ou informação – porque faz com que esses crimes sejam rentáveis e isso encoraja aqueles que os cometem. Por outro lado, e em especial no caso de sequestro e extorsão, a pessoa lesada é compelida a cooperar divergindo, portanto, do comportamento esperado[40].

Assim como a pressão reputacional, para funcionar, a pressão institucional demanda consequências. A diferença é que, enquanto as consequências da reputação são informais, as consequências institucionais são formais, codificadas e tangíveis. Essas consequências podem ser de natureza punitiva, i.e., sanções, ou recompensável, i.e., incentivos. A ideia por trás dessas recompensas é formalizar coerção. Mesmo as leis proibitivas têm esse caráter: ao passo que, tanto prescrevem quanto punem.

As leis são tão eficazes quanto a capacidade da sociedade em aplicá-las. Não é o suficiente editar uma lei, se você não é capaz de cercear os infratores; ou, as leis não passarão de um mero impedimento. Isso é particularmente visível no Brasil onde a expressão *pegar* é aplicada às leis que não são efetivamente cumpridas pelo poder público, apesar de sua existência[41].

As sanções são classificadas em três categorias básicas: confisco de recursos ou bens (e.g., multas e trabalhos forçados), humilhação (e.g., registro de criminosos sexuais) e castigos físicos (e.g., encarceramento, danos físicos e execução). Castigos físicos e humilhações foram as formas mais prevalentes durante a história da humanidade.

---

[39] ibid.
[40] ibid.
[41] Vide (FERNANDES, 2016)

A maioria das sanções atuais consiste no encarceramento ou sanções pecuniárias. O encarceramento remove o criminoso da sociedade por um período de tempo, e impede-o de cometer novos delitos. As sanções pecuniárias podem ser difíceis de implantar e, portanto, dignas de uma discussão mais longa.

É vital para essas sanções serem altas o suficiente para tornar o comportamento que se quer impedir, inviável; do contrário sua eficácia é diminuída pelo poder econômico do infrator. Por esse motivo, a maioria das multas incidentes sobre impostos devidos, é calculada com base percentual, incidente sobre o valor devido, e não sobre valores fixos e diferenciados para as diversas categorias de devedores (IGNACIO, 2012). Outro exemplo é o sistema de pontos, adotado pelo Departamento Estadual de Trânsito (DETRAN), junto ao sistema de multas. Impostos são outras formas de pressões institucionais. Na verdade, não proíbem, mas objetivam a redução no escopo da evasão.

O contraponto às sanções são os incentivos, i.e., gratificar alguém por cooperar, e.g., deduções fiscais, reembolso fiscal mais rápido, bônus aos funcionários, certificações etc. O problema da remuneração institucional recai essencialmente sobre seu custo. É bom ressaltar que incentivos e penalidades financeiras interagem estranhamente com outras categorias de pressões sociais. A simples existência de regras ou leis podem contrariar questões existentes de reputação e moral.

Vale ressaltar ainda, que, reputação negativa também pode ser institucionalizada, e.g., listas negras ou apartheid. Leis também formalizam compromisso, e.g., contratos.

Em suma, as leis formalizam as normas sociais que a reputação tradicionalmente compele. Todas estas pressões institucionais permitem à expansão da reputação, dando às pessoas um sistema para confiarem, libertando-as da necessidade de confiar em cada indivíduo.

Em certas ocasiões e para algumas pessoas, a lei não é impedimento suficiente e o risco vale a pena. Noutras, a própria legislação traz consigo meios, defeitos ou incentivos. Existem várias maneiras pelas quais as pressões institucionais falham (REZENDE, 2002).

Excesso ou Falta – vimos como a pressão institucional é necessária para aumentar as pressões morais e reputacionais nas sociedades mais complexas. No entanto, mais pressão institucional nem sempre acarreta melhoria.

Efeito contrário – as leis podem nem sempre ter o efeito desejado. Um aumento na severidade da pena, nem sempre se traduz em uma queda da criminalidade, um aumento na probabilidade de castigo muitas vezes sim. Muitas vezes é mais importante atacar as causas sociais da criminalidade que modificar a lei.

Aplicabilidade – nem sempre é possível aplicar determinada lei. O Direito Internacional só importa na medida em que os países estão dispostos a respeitá-lo e são capazes de aplicá-lo. Lavagem de dinheiro e crimes cibernéticos figuram entre os principais; mas aqui, se incluem problemas como tráfico de drogas, armas e até animais e espécies em extinção.

Brechas[42] – algumas leis podem conter "brechas". Isso pode acontecer por erro (e.g., incapacidade de antecipar algum novo desenvolvimento tecnológico) ou acontecer de forma deliberada (e.g., má fé). Como exemplo, podemos citar praticamente toda forma de elisão fiscal, incluindo formas sofisticadas como o Double Irish (THE NEW YORK TIMES, 2012). Geralmente, as brechas são subversões das pressões institucionais usadas para fins que não idealizados originalmente, e.g., o uso da lei do direito autoral, que originalmente fora criada para salvaguardar os autores, como base para tentativas governamentais de controle da Internet e da população na Austrália (LIA, 2012).

Inconsistência – certas leis podem ser aplicadas de forma inconsistente. Se as leis não são objetivas, universais, e universalmente aplicadas, elas são vistas como desleais, e a injustiça pode gerar efeitos nefastos.

Variabilidade – as leis não se aplicarem igualmente a todo tipo de infrator. Podemos agrupar os infratores em três grupos. O primeiro é formado pelas pessoas que conhecem a lei, são indiferentes a sua corretude, e escolhem infringi-la conscientemente, e.g., ladrões, assassinos, sequestradores etc. O

[42] Lacuna ou oportunidade na lei (BRASIL, 2022).

segundo, são as pessoas que conhecem a lei, acreditam que a lei está errada, e escolhem infringi-la, e.g., usuários de droga e alguns piratas cibernéticos. Há também uma terceira categoria: são aquelas pessoas que não sabem que estão infringindo a lei, ou não percebem como suas ações afetam o grupo, e.g., incorrer em certas práticas sexuais tidas como ilegais (CNN, 2009; BLODGET, 2010).

4.5.1 Ilustração: Algoritmos de justiça criminal

A inteligência artificial é amplamente utilizada pelo sistema de justiça (CGTI/ACS, 2019) na esperança de que a tecnologia traga mais celeridade nos processos, ajude a mitigar falhas e a construir uma sociedade mais justa. Para a população em geral, o primeiro benefício é a expectativa de ter uma justiça mais eficiente, ágil e com maior grau de acerto. Também se espera que fique muito mais simples obter o serviço de representação jurídica.

A inteligência artificial no direito também afeta os serviços dos órgãos do Poder Judiciário. Vale citar o caso da Estônia, que está desenvolvendo um “juiz robô” para analisar disputas legais simples, i.e., menos de € 7.000,00 (NIILER, 2019). O protocolo é simples, as duas partes enviam os documentos relevantes para o caso e a inteligência artificial toma uma decisão, que pode ser revista por um juiz. Por enquanto, as autoridades estonianas gostam da ideia de uma IA resolvendo disputas simples, deixando mais tempo para juízes e advogados humanos resolverem os casos mais difíceis.

Mas nem todas as incursões da IA na seara jurídica são oriundas do poder público. As *lawtechs*, ou *legaltechs*, são empresas com matriz tecnológica que prestam um serviço ou desenvolvem um produto de base tecnológica na área jurídica. Entre as soluções elaboradas por elas destacamos: plataformas de automação; gestão de documentos; conformidade; comunicação; mediação; gestão jurídica; monitoramento de dados públicos; redes profissionais; resolução de conflitos online etc. (TOEWS, 2019). Todavia, foram os avaliadores de risco que ganharam notoriedade recente na mídia por maleficiar minorias (ISRANI, 2017; YONG, 2018).

Os algoritmos de justiça criminal, também conhecidos como "avaliadores de risco" ou "métodos baseados em evidências", são ferramentas que se supõe sejam capazes de prever o comportamento futuro de réus e pessoas encarceradas. As metodologias variam, mas essencialmente, trabalham em cima das probabilidades de o réu reincidir antes do julgamento e deixar de comparecer ao julgamento.

Esses algoritmos são usitados para definir fiança, determinar sentenças e até auxiliam magistrados a deliberar sobre a culpa ou inocência do réu. Observa-se, no entanto, que a maioria desses métodos possui pouca ou nenhuma transparência (PROPUBLICA, 2016). Após exame minucioso por especialistas em justiça criminal, diversas denúncias sobre a opacidade, controvérsias e inconstitucionalidade dessas ferramentas surgiram na mídia (ISRANI, 2017; YONG, 2018).

Muitos algoritmos analisam características pessoais como idade, sexo, geografia (i.e., local de nascimento, habitação e trabalho), histórico familiar e status de emprego. Como decorrência duas pessoas acusadas do mesmo crime podem receber decisões, e.g., de fiança ou sentença, nitidamente diferentes a depender das informações prestadas. Sem contar que por apresentarem pouca ou nenhuma transparência, ou regras comprovadamente justas, fica inviável avaliar ou contestar as decisões.

Os três principais sistemas usados, em sua totalidade ou adaptados, atualmente são: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) criado pela Northpointe, o Public Safety Assessment (PSA), desenvolvido pela Fundação Laura e John Arnold, e Level of Service Inventory Revised (LSI-R), desenvolvido pela empresa canadense Multi-Health Systems.

4.5.1.1 LSI-R

O LSI-R é um instrumento de avaliação de risco, parte da investigação de présentença, realizada sobre réu para informar decisões correcionais de custódia, supervisão e provisão de serviços. Essas avaliações ambicionam determinar o risco que um réu representa para a sociedade por meio da reincidência, i.e., cometendo outro crime, e o tratamento para reabilitar o réu.

O LSI-R mede 54 fatores de risco e necessidades em 10 domínios criminogênicos. Os domínios preditores medidos são: histórico criminal, educação / emprego, situação financeira, relações familiares / conjugais, acomodação, lazer e recreação, companhias, uso de álcool ou drogas, saúde emocional / mental e atitudes e orientações (ANDREWS e BONTA, 1995). O teste não se concentra apenas em cada domínio individualmente, mas também em como eles interagem.

O teste é administrado por meio de uma entrevista estruturada entre o avaliador e o réu, com a recomendação de que documentação extra seja coletada junto a familiares, empregadores, arquivos do caso, testes de drogas e outras fontes relevantes, conforme necessário (ANDREWS e BONTA, 1995).

A pontuação total produzida é um indicativo do número de itens preditores (num total de 54) apresentados pelo réu. O escore LSI-R é então atuarialmente associado a uma probabilidade de reincidência, derivada das taxas de reincidência observadas em infratores previamente avaliados. Por fim, as pontuações de domínio são usadas para identificar os tratamentos mais promissores para o réu. (ANDREWS e BONTA, 2003).

Um conjunto significativo de estudos comprovam a validade preditiva do LSI-R para: infrações institucionais (BONTA, 1989), transgressão de liberdade provisória (ANDREWS, KIESSLING e MICKUS, 1986), violações de liberdade condicional (BONTA e MOTIUK, 1990) e preditor de futuras ofensas (ANDREWS e BONTA, 1995; GOGGIN, GENDREAU e GRAY, 1998).

(WRIGHT, CLEAR e DICKSON, 1984) testaram a validade preditiva do modelo usado pelo estado de Wisconsin em amostras obtidas nos estados de Nova York e Ohio. Embora o modelo de Wisconsin tenha demonstrado validade preditiva para a amostra em que foi criada (BAIRD, HINES e BEMUS, 1979), ele não demonstrou validade preditiva para as amostras de Nova York ou Ohio (WRIGHT, CLEAR e DICKSON, 1984). Essa descoberta serve como um lembrete para que os sistemas de classificação devam ser validados para suas respectivas populações.

Como o LSI-R representa uma medida padronizada de riscos e necessidades criminogênicas, atuarial, fundamentada teoricamente e empiricamente embasada,

possui um potencial considerável para automação, melhorando a alocação de recursos e a qualidade do serviço prestado como um todo.

As avaliações de risco/necessidades, com base estatística, foram aceitas como métodos corrente e válidos para organizar grande parte das informações críticas relevantes para gestão de infratores em ambientes correcionais (QUINSEY, HARRIS, *et al.*, 1998). (GROVE, ZALD, *et al.*, 2000) (SWETS, DAWES e MONAHAN, 2000) concluíram que avaliações estatísticas são superiores ao julgamento humano.

4.5.1.2 COMPAS

COMPAS é uma ferramenta de gerenciamento e suporte à decisão, com base estatística, desenvolvido para avaliar a probabilidade de um réu reincidir. Compreende uma total de 43 escalas, incluindo 4 escalas de alta ordem que usam itens de vários domínios e 17 escalas de avaliação de risco e necessidades das mulheres.

O COMPAS foi desenvolvido para ser configurável pelo operador a fim de atender as idiossincrasias locais do sistema judiciário e adequação a diferentes populações. Após completar uma consulta no COMPAS, o operador interpreta o gráfico de barras que exibe as pontuações de cada área e poderá interpretar a tipologia associada.

A interpretação precisa de um gráfico de barras do COMPAS exige que o operador leve em consideração todos as escalas de pontuação alta. As teorias criminológicas fornecem uma estrutura para ajudar a entender a interrelação entre as diferentes escalas. Dessa forma, existem diferentes níveis de interpretação.

No nível elementar, os fatores são abordados isoladamente, desconsiderando a influência que os demais resultados podem exercer sobre o fator analisado. No nível intermediário, são identificados os fatores criminogênicos que estão interrelacionados, i.e., inicia o processo de análise das áreas de necessidade que se sugestionam. No último nível, a interpretação é totalmente integrada, i.e., emprega teo-

rias criminológicas para explicar padrões de comportamento criminoso, e ajuda os operadores a entender possíveis causas subjacentes.

Como vimos anteriormente, os sistemas de classificação devem ser validados para suas respectivas populações e quando esse dogma é desrespeitado as consequências podem ser nefastas.

Em estudo publicado em maio de 2016, com os dados de mais de 7.000 pessoas avaliadas na fase pré-julgamento, a ProPublica descobriu que os réus negros, analisados pelo COMPAS, eram muito mais propensos do que os réus brancos a serem incorretamente considerados com maior risco de reincidência, enquanto os réus brancos eram mais propensos do que os réus negros a serem sinalizados incorretamente como de baixo risco (PROPUBLICA, 2016).

O estudo da ProPublica (PROPUBLICA, 2016), acerca do COMPAS, e outros estudos mais amplos (SKEEM e LOWENKAMP, 2016), nos remetem às 20 recomendações de (FLORIDI, COWLS, *et al.*, 2018), especialmente as categorias normativas de avaliação e desenvolvimento. Como veremos a seguir, o PSA é um terceiro sistema, i.e., ferramenta de avaliação de risco, que recentemente “abriu” e “explicou” o seu funcionamento para o público.

4.5.1.3 PSA

O PSA é uma avaliação baseada em pesquisa que assessora juízes a tomar decisões mais esclarecidas na fase de pré-julgamento. Ele usa nove fatores para calcular a probabilidade de sucesso de uma pessoa que aguarda liberação durante o pré-julgamento. Os fatores incluem a idade atual da pessoa, condenações anteriores, acusações pendentes e o não comparecimento em julgamentos anteriores (LAURA AND JOHN ARNOLD FOUNDATION, 2020).

Após análise desses fatores, o PSA gera duas pontuações distintas (que variam de 1 a 6, com pontuações mais altas indicando um maior nível de risco): risco de incorrer em nova atividade criminosa, New Criminal Activity (NCA), e falta de comparência em tribunal, Failure to Appear (FTA). A avaliação pode gerar um alerta, New

Violent Criminal Activity (NVCA), para indicar se a pessoa apresenta uma alta probabilidade de ser presa por um novo crime violento caso seja libertada durante o período pré-julgamento.

O PSA não substitui o juiz nem compromete a discrição ou autoridade do magistrado. A decisão sobre liberar ou deter um réu sempre recai sobre o juiz. Para ajudar os tomadores de decisão a usarem as pontuações do PSA em suas decisões, são criadas, em parceria com as instituições locais, uma estrutura de decisão e uma matriz de condições de liberação, que refletem os estatutos locais, regras dos tribunais e preferências políticas.

A fim de mitigar questões de opacidade, controvérsias e inconstitucionalidade, são publicizadas a estrutura de decisão e a matriz de condições de liberação. Além disso, recentemente, o fabricante disponibilizou vasta, e acessível, documentação sobre o design, implementação, funcionamento, teste e auditoria do sistema e da metodologia que é seguida para a adaptação do sistema as diversas realidades locais (LAURA AND JOHN ARNOLD FOUNDATION, 2020).

A essa altura o leitor deve se indagar sobre a real efetividade da implantação dos princípios éticos estabelecidos - materializados nos diversos relatórios, manuais e códigos de ética para IA - na cadeia de produção dos sistemas de IA. Tudo aponta para a necessidade de uma legislação específica para a IA - i.e., que aspire considerações éticas, legais e sociais - e instituições que assegurem sua devida implantação, execução e manutenção, e.g., Regulamento do Parlamento Europeu e do Conselho que estabelece regras harmonizadas em matéria de inteligência artificial (regulamento inteligência artificial) e altera determinados atos legislativos da união (EUROPEAN PARLIAMENT, 2021).

Como veremos a seguir, políticos têm sustentado a tese de que pressões sociais informais e normativas não foram suficientes para assegurar uma IA ética, i.e., que respeita os direitos fundamentais, não discrimina, segura, justa, auditável, transparente e imparcial.

4.6 POLÍTICAS

Em setembro de 2019 o senador Styvenson Valetim (PODEMOS-RN) publicou o Projeto de Lei nº 5.051/2019, que, em seu art. 1º, “estabelece os princípios para o uso da Inteligência Artificial no Brasil” (BRASIL, 2019). Composto por 7 artigos, o projeto aborda: os princípios que devem nortear o desenvolvimento da inteligência artificial, art. 2 e 3; como a IA precisa estar condicionada a ser uma apoiadora da tomada de decisão humana, a proporcionalidade dessa supervisão e a responsabilização civil do supervisor, art. 4; e as diretrizes e princípios para a atuação da união, dos estados, do Distrito Federal e dos municípios no desenvolvimento da IA no Brasil, art. 5 e 6.

Infelizmente, emanam inconsistências inconciliáveis do projeto de lei (PL). Numa primeira leitura, tem-se a impressão de que o projeto busca desacelerar a evolução tecnológica no setor de IA, e.g., o artigo 4: “Os sistemas decisórios baseados em Inteligência Artificial serão, sempre, auxiliares à tomada de decisão humana” (BRASIL, 2019), o que vai de encontro à justificativa que consta na proposta, “como se observa, não se trata de frear o avanço da tecnologia, mas de assegurar que esse desenvolvimento ocorra de modo harmônico com a valorização do trabalho humano, a fim de promover o bem-estar de todos” (BRASIL, 2019).

O projeto de lei cria a figura do supervisor que deverá ser humano e civilmente responsável pelas decisões tomadas pela IA. O cargo se assemelha ao de operador na lei geral de proteção de dados (LGPD), mas sem a possibilidade de ser pessoa jurídica, de direito público ou privado, segundo o artigo 5º parágrafo 7 (BRASIL, 2019). Apesar de garantir para o supervisor exigibilidade compatível (com o tipo, a gravidade e as implicações da decisão submetida), a responsabilização civil por danos decorrentes da IA será de exclusividade do supervisor, não havendo solidariedade prevista para o criador/mantenedor da IA. Em 8 de dezembro de 2022, o PL foi juntado ao Relatório Final da Comissão de Juristas, instituída pelo Ato do Presidente do Senado nº 4, de 2022 (Senado Federal, 2022).

4.6.1 EBIA

No âmbito da Estratégia Brasileira para a Transformação Digital (E-Digital)[43], em julho de 2021 foi apresentada a Estratégia Brasileira de Inteligência Artificial (EBIA). A estratégia propõe-se a nortear as ações do estado brasileiro em prol do desenvolvimento de ações que estimulem a pesquisa, inovação e desenvolvimento de IA conscientes, éticas e em prol de um futuro melhor (MCTIC, 2021).

Calcada em recentes recomendações da OCDE sobre Inteligência Artificial[44], a tática brasílica estabelece nove eixos temáticos, a seguir identificados. Eixos transversais: Legislação regulação e uso ético, Governança de IA e Aspectos internacionais. Eixos verticais: Educação, Força de Trabalho e Capacitação, PD&I e Empreendedorismo, Aplicação nos setores Produtivos, Aplicação no Poder Público e Segurança Pública.

Espera-se que a Inteligência Artificial transforme profundamente as estruturas econômicas e sociais do país. Portanto, é fundamental que o Governo brasileiro estabeleça políticas públicas para endereçar tais mudanças, abrangendo não apenas a tecnologia e a indústria, mas também a educação, o emprego e o bem-estar. (MCTIC, 2021)

Inicialmente, a EBIA apresenta seis objetivos estratégicos que levam em consideração todo o ecossistema tecnológico, reforçando a importância dessa iniciativa para a evolução da gestão pública, e que poderão posteriormente ser desdobrados em ações específicas. São eles:

(EB1) Contribuir para a elaboração de princípios éticos para o desenvolvimento e uso de IA responsáveis;

(EB2) Promover investimentos sustentados em pesquisa e desenvolvimento em IA;

---

[43] Aprovada em março de 2018, pelo Decreto nº 9.319/2018 e pela Portaria MCTIC nº 1.556/2018
[44] Vide (OECD, 2019).

(EB3) Remover barreiras à inovação em IA;

(EB4) Capacitar e formar profissionais para o ecossistema da IA;

(EB5) Estimular a inovação e o desenvolvimento da IA brasileira em ambiente internacional;

(EB6) Promover ambiente de cooperação entre os entes públicos e privados, a indústria e os centros de pesquisas para o desenvolvimento da Inteligência Artificial.

A vista disso, a EBIA guia as políticas públicas para desenvolver a IA Ética. O objetivo é abordar os aspectos probos para beneficiar as pessoas com uma abordagem regulatória equilibrada, propiciar o crescimento inclusivo e segurança jurídica com foco em princípios éticos e respeito aos direitos humanos.

Além da EBIA, ao redor do mundo há ações estratégicas para implementar e promover a Inteligência Artificial. Exempli gratia, a Declaração de Toronto 2018[45], o AI Framework de Singapura[46] e a Resoluções do Parlamento Europeu[47] voltados à Inteligência Artificial.

4.6.2 Algorithmic Accountability Act

Numa verve mais positivista, o *Algorithmic Accountability Act of 2019* (AAA19), inspirado nos princípios de vários artigos da General Data Protection Regulation (GDPR), define o papel, direitos e deveres da autoridade de proteção de dados, submetido à Federal Trade Commission (FTC). O PL se encarrega de discriminar os atuais sistemas, caracterizar a elegibilidade das entidades responsáveis, i.e., pessoa, parceria ou corporação, além das exigibilidades, i.e., estudos, a serem fornecidas pelos entes.

O PL estadunidense, ao contrário do brasileiro, distingue os sistemas em: sistema de informação (SI) – cabendo ainda a possibilidade de ser ou não de alto risco

[45] Vide (AMNESTY INTERNATIONAL, 2018)
[46] Vide (PDPC, 2020).
[47] Vide (EUROPEAN PARLIAMENT, 2021).

– e sistema de decisão automatizada (SDA) – cabendo ainda a possibilidade de ser ou não de alto risco.

Um sistema de informação é um processo automatizado, ou não, que envolve informações pessoais – como coleta, registro, organização, estruturação, armazenamento, alteração, recuperação, consulta, uso, compartilhamento, divulgação, disseminação, combinação, restrição, apagamento ou destruição de dados pessoais em formação – e não inclui sistemas de decisão automatizados (UNITED STATES, 2019).

É tido de alto risco quando: levando-se em conta a novidade da tecnologia utilizada e a natureza, escopo, contexto e finalidade do sistema de informação, representa um risco significativo para a privacidade ou segurança das informações pessoais dos consumidores; envolve informações pessoais de um número significativo de consumidores sobre raça, cor, nacionalidade, opiniões políticas, religião, associação sindical, dados genéticos, dados biométricos, saúde, gênero, identidade de gênero, sexualidade, orientação sexual, condenações criminais, ou prisões; monitora sistematicamente um grande local físico acessível ao público; ou atende a qualquer outro critério estabelecido pela FTC (UNITED STATES, 2019).

Um sistema de decisão automatizada é um processo computacional – incluindo os derivados de aprendizado de máquina, estatística ou outras técnicas de processamento de dados ou inteligência artificial – que toma uma decisão ou facilita a tomada de decisão humana, que afeta os consumidores, i.e., indivíduos (UNITED STATES, 2019).

É tido de alto risco quando atende aos seguintes critérios: levando-se em consideração a novidade da tecnologia empregada e a natureza, escopo, contexto e objetivo do sistema, apresenta um risco significativo: a privacidade ou segurança das informações pessoais dos consumidores; ou resultar ou contribuir para decisões imprecisas, injustas, tendenciosas ou discriminatórias que afetam os consumidores;

Toma decisões ou facilita a tomada de decisões humanas, com base em avaliações sistemáticas e extensas dos consumidores, incluindo tentativas de analisar ou prever aspectos sensíveis de suas vidas, como desempenho no trabalho, situação econômica, saúde, preferências pessoais, interesses, comportamento, localiza-

ção ou movimentos, que - alteram os direitos legais dos consumidores ou impactam significativamente os consumidores;

Envolve as informações pessoais de um número significativo de consumidores sobre raça, cor, nacionalidade, opiniões políticas, religião, associação sindical, dados genéticos, dados biométricos, saúde, gênero, identidade de gênero, sexualidade, orientação sexual, condenações criminais, ou prisões; monitora sistematicamente um local físico, amplo e acessível ao público; ou atende a qualquer outro critério estabelecido pela FTC (UNITED STATES, 2019).

O PL se aplica a qualquer pessoa, parceria ou corporação sobre a qual a FTC tenha jurisdição que:

> tenha mais de US$ 50.000.000,00 de receitas brutas anuais médias para o período de três anos tributáveis anterior ao ano fiscal mais recente;
>
> possua ou controle informações pessoais de mais de 1.000.000 de consumidores ou 1.000.000 de dispositivos de consumidores;
>
> possua, seja operado ou controlado substancialmente por uma pessoa, parceria ou corporação; ou
>
> seja um corretor de dados ou outra entidade comercial que, como parte substancial de seus negócios, coleta, reúne ou mantém informações pessoais sobre um indivíduo que não é um cliente ou funcionário dessa entidade para vender ou negociar as informações ou fornecer acesso a terceiros.

O AAA19 orienta a FTC a exigir das entidades – que usam, armazenam ou compartilham informações pessoais – a realização da avaliação de impacto do sistema de decisão automatizada e avaliação de impacto na proteção de dados (UNITED STATES, 2019).

Para a referida lei, a avaliação de impacto na proteção de dados, significa um estudo que avalia até que ponto um sistema de informação protege a privacidade e a segurança das informações pessoais que o sistema processa.

Já a avaliação de impacto do sistema de decisão automatizada consiste em um estudo que avalia o sistema de decisão automatizada e o processo de desenvolvimento do sistema, incluindo design e dados de treinamento do modelo, quanto a

impactos na precisão, justiça, preconceito, discriminação, privacidade e segurança. Deve incluir, mas não se limita a:

> (k1) uma descrição detalhada do sistema, i.e., seu design, treinamento, dados e finalidade;
>
> (k2) uma avaliação dos benefícios e custos relativos ao sistema, à luz de sua finalidade, levando em consideração fatores relevantes, incluindo:
>
> > (i) práticas de minimização de dados;
> >
> > (ii) a duração pela qual as informações pessoais e os resultados do sistema são armazenados;
> >
> > (iii) quais informações sobre o sistema estão disponíveis para os consumidores;
> >
> > (iv) até que ponto os consumidores têm acesso aos resultados do sistema e podem corrigir ou objetar seus resultados; e
> >
> > (v) os destinatários dos resultados do sistema;
>
> (k3) uma avaliação dos riscos impostos pelo sistema à privacidade ou segurança das informações pessoais dos consumidores e aos riscos que o sistema pode resultar ou contribuir para decisões imprecisas, injustas, tendenciosas ou discriminatórias que afetam os consumidores; e
>
> (k4) as medidas que a entidade coberta empregará para minimizar os riscos descritos no subparágrafo (k3), incluindo salvaguardas tecnológicas e físicas.

A exigência da realização, tanto para sistemas novos quanto existentes, das avaliações de impacto do sistema de decisão e na proteção de dados será mandatória após 2 anos de promulgada a lei (UNITED STATES, 2019). O AAA19 enfatiza, mas não obriga, a importância desses estudos serem realizados em consulta com terceiros, incluindo auditores e especialistas em tecnologia independentes. Já os documentos, podem ser tornados público pela entidade a seu exclusivo critério (UNITED STATES, 2019).

Em 3 de Fevereiro de 2022 o PL AAA19 sofreu uma série de modificações e foi rebatizado como *Algorithmic Accountability Act of 2022* (AAA22). A expansão do

projeto de lei, de 15 para 50 páginas (entre as versões 2019 e 2022), facultou definições adequadas de termos como sistemas de decisão automatizados (SDA) e 'entidade coberta' (i.e., pessoa ou organização à qual o projeto de lei se aplica).

De acordo com a versão preliminar atual do AAA22, Sistemas de Decisão Automatizados de Alto Risco (SDAAR) incluem aqueles que podem contribuir para imprecisão, viés ou discriminação; ou facilitar a tomada de decisões sobre aspectos sensíveis da vida dos consumidores via avaliação comportamental. Além disso, um SDA, ou SI envolvendo dados pessoais, é considerado de alto risco se (UNITED STATES, 2022):

(1) Levanta questões de segurança ou privacidade;

(2) Envolve as informações pessoais de um número significativo de pessoas; ou

(3) Monitora sistematicamente um grande local físico acessível ao público.

Outrossim, o PL estabelece os requisitos para a avaliação de impacto, que devem ser realizadas por entidades abrangidas, em consulta com as partes interessadas internas e externas, para SDA e SDAAR e devem (UNITED STATES, 2022):

(κ1) Incluir uma descrição do processo existente para a mesma decisão juntamente com uma análise comparativa dos benefícios, necessidade e uso pretendido;

(κ2) Identificar e descrever as consultas às partes interessadas, bem como as recomendações recebidas;

(κ3) Realizar testes e avaliações contínuas para riscos de privacidade e medidas para aumentar a privacidade;

(κ4) Documentar métodos, métricas, conjuntos de dados adequados, critérios para desempenho bem-sucedido e outros padrões;

(κ5) Realizar testes contínuos e avaliação de desempenho em condições de teste e implantadas e para diferentes grupos demográficos;

(κ6) Apoiar e realizar treinamento para agentes relevantes sobre os riscos e melhores práticas para SDA similares;

(κ7) Avaliar a necessidade de limitações no uso de SDA e desenvolver tais limitações no produto ou em seus termos de uso;

(κ8) Manter a documentação dos metadados dos conjuntos de dados e outras informações de entrada, usadas no desenvolvimento, teste e manutenção do SDA, bem como as razões para usar tais dados e alternativas exploradas;

(κ9) Avaliar os direitos dos consumidores em termos de transparência, com um aviso claro sobre o uso de SDA e informações sobre destinatários de decisões de terceiros; avaliações contrafactuais; e mecanismos para contestar, corrigir, apelar e recusar;

(κ10) Identificar prováveis efeitos adversos de forma estruturada e avaliar estratégias de mitigação relevantes;

(κ11) Descrever a documentação dos processos de desenvolvimento, teste e implantação;

(κ12) Identificar capacidades, ferramentas, padrões e protocolos para melhorar o SDA ou a avaliação de impacto em áreas como desempenho (e.g., precisão, robustez, confiabilidade), imparcialidade (i.e., viés, não discriminação), transparência, explicabilidade, contestabilidade, uma oportunidade de recurso, privacidade e segurança, segurança pessoal e pública e outras áreas;

(κ13) Incluir uma justificativa se algum dos requisitos acima não for seguido; e

(κ14) Realizar e documentar outros estudos e avaliações que a FTC considere apropriados.

Os relatórios resumidos devem ser submetidos à FTC anualmente. Para novos SDA e SDAAR, também é necessário um relatório resumido inicial antes da implantação. Os relatórios resumidos precisam conter informações apenas sobre o seguinte (UNITED STATES, 2022):

(K1) Descrição da decisão crítica;

(K2) Finalidade pretendida da SDA;

(K3) Partes interessadas consultadas;

(K4) Métodos, métricas e o que é considerado desempenho bem-sucedido;

(K5) Resultados de avaliações de desempenho em diferentes grupos demográficos;

(K6) Limitações declaradas publicamente sobre o uso do SDA;

(K7) Se e como as medidas de transparência ou explicabilidade são implementadas; e

(K8) provável impacto negativo material e respectivas estratégias de mitigação.

Além de exigir que empresas e organizações conduzam análises de impacto, o AAA22 estabelece um repositório público dentro da FTC. Esse armazém visa prover um certo nível de transparência e responsabilidade entre o produtor da IA e os clientes ao armazenar, de forma segura, os meios pelos quais um determinado sistema de IA chegou a uma determinada conclusão (UNITED STATES, 2022).

Ainda assim, a AAA22 delega muitas escolhas importantes para à FTC, o que não é necessariamente negativo[48], mas a falta de especificidade do PL é um problema nos casos em que é desnecessariamente vago em delegar esses detalhes à FTC.

4.6.3 Ethics guidelines for trustworthy AI

A Comissão Europeia publicou suas diretrizes para o desenvolvimento e a implementação de padrões éticos da Inteligência Artificial. Intitulado "Ethics guidelines for trustworthy AI", o texto foi concebido como um ponto de partida para o debate sobre o tema.

As orientações contidas nele, visam promover a investigação, a reflexão e o debate, sobre um quadro ético para a IA, a nível mundial. Apesar de não ser vinculante, ele serve como um arcabouço para orientar ações futuras.

O documento estabelece que uma IA de confiança tem três componentes, que devem ser observados ao longo de todo o ciclo de vida do sistema:

---

[48] A maioria das leis aprovadas pelo congresso norte americano delega alguma autoridade às agências executivas (CLOUSER MCCANN e SHIPAN, 2022).

(1) **IA Legal** - garante a conformidade com a legislação e regulamentação aplicáveis;

(2) **IA Ética** - assegura a observância de princípios e valores éticos; e

(3) **IA Robusta** – incumbe-se, tanto do ponto de vista técnico como do ponto de vista social, que a IA não cause danos intencionais.

Cada um destes componentes é necessário, mas não suficiente *per se*. O ideal é que funcionem em consonância. A IA Legal não é coberta pelo documento, pois encontra-se em construção, mas estará em sintonia com as várias regras juridicamente vinculantes em vigor, a nível europeu, nacional e internacional, que são aplicáveis ou relevantes para o desenvolvimento, implantação e utilização da IA.

O capítulo I trata das bases da IA de confiança. Identifica os princípios éticos e respectivos valores que devem ser respeitados durante o desenvolvimento, implantação e utilização da IA (i.e., desenvolver, implantar e utilizar os sistemas de IA de uma forma consentânea com os princípios éticos de: respeito da autonomia humana, prevenção de danos, equidade e explicabilidade) (HIGH-LEVEL EXPERT GROUP ON AI, 2019).

(1) **O princípio do respeito da autonomia humana** - Os direitos fundamentais que visam garantir o respeito da liberdade e da autonomia dos seres humanos.

(2) **O princípio da prevenção de danos** - A IA não deve causar danos, ou agravá-los, nem afetar negativamente os seres humanos de qualquer outra forma.

(3) **O princípio da equidade** - O desenvolvimento, a implantação e a utilização da IA devem ser equitativos.

(4) **O princípio da explicabilidade** - Significa que os processos são transparentes, as capacidades e a finalidade da IA são publicizadas e as decisões explicáveis aos que são afetados de forma direta e indireta.

A fim de reconhecer e sanar eventuais conflitos entre os princípios, sugere-se prestar especial atenção a situações que envolvam grupos mais vulneráveis, e.g., crianças, pessoas com deficiência e outros grupos historicamente desfavorecidos ou

em risco de exclusão, e situações caracterizadas por assimetrias de poder ou de informação, e.g., relações patronais ou relações de consumo.

Também, deve-se reconhecer a suscetibilidade da IA a impactos negativos, de difícil previsibilidade, detecção e mensuração, e.g., na democracia, no Estado de direito e na justiça distributiva, ou na própria mente humana; e a adoção de medidas adequadas para atenuar os riscos, proporcionalmente, quando necessário.

O capítulo 2 trata da concretização da IA de confiança. Fornece orientações sobre como alcançar a IA de confiança através de sete requisitos que devem ser cumpridos. Pondera sobre os métodos técnicos e não técnicos para assegurar a aplicação desses requisitos. (HIGH-LEVEL EXPERT GROUP ON AI, 2019).

Os requisitos a seguir, procuram assegurar que o desenvolvimento, a implantação e a utilização da IA satisfaçam as exigências para uma IA de confiança. A saber:

(r1) **ação e supervisão humanas** – inclui os direitos fundamentais, a ação humana e a supervisão humana;

(r2) **solidez técnica e segurança** – inclui a resiliência perante ataques e a segurança, os planos de recurso e a segurança geral, a exatidão, a fiabilidade e a reprodutibilidade;

(r3) **privacidade e governação dos dados** – inclui o respeito à privacidade, a qualidade e a integridade dos dados e o acesso aos dados;

(r4) **transparência** – inclui a rastreabilidade, a explicabilidade e a comunicação;

(r5) **diversidade, não discriminação e equidade** – inclui a prevenção de enviesamentos injustos, a participação das partes interessadas, e o conceito de acessibilidade e a concessão universal;

(r6) **bem-estar ambiental e societal** – inclui a sustentabilidade e o respeito ao meio ambiente, o impacto social, a sociedade e a democracia;

(r7) **responsabilização** – inclui a auditabilidade, a minimização e a comunicação dos impactos negativos, as soluções de compromisso e as vias de recurso (HIGH-LEVEL EXPERT GROUP ON AI, 2019).

Destarte, estes requisitos objetivam, pela perspectiva técnica, facilitar a rastreabilidade e a auditabilidade da IA, sobretudo em contextos ou situações críticas; conscientizar da possível existência de conflitos fundamentais entre os diferentes princípios e requisitos; promover a investigação e a inovação para ajudar a avaliar a IA e melhorar o cumprimento dos requisitos; além de identificar, avaliar, documentar e comunicar continuamente essas soluções.

Ambiciona-se no panorama social, ser transparente sobre o fato das pessoas estarem lidando com uma IA; comunicar, de forma clara e proativa, às partes interessadas, informações sobre as capacidades e as limitações da IA, i.e., permitindo-lhes criar expectativas realistas, e sobre a forma como os requisitos são aplicados; e assim, informar ao público todo o ciclo de vida da IA;

Ao promover a formação e a educação, o público passa a ter mais conhecimento e formação sobre a IA de confiança. E, ao divulgar os resultados e as questões em aberto, almejamos a formação de uma nova geração de peritos em IA ética.

Estes requisitos – juntamente com direitos, princípios e valores previamente arrolados – são introduzidos no ciclo de desenvolvimento da IA por meio dos métodos técnicos e não técnicos. Uma vez que esses sistemas inteligentes evoluem continuamente, a concretização da IA de confiança é um processo contínuo, implementado por métodos técnicos e não técnicos.

Métodos técnicos - são os métodos que permeiam as fases de design, implementação e utilização do sistema inteligente para se auferir uma IA de confiança. Os métodos apresentados são complementares ou alternativos entre si, uma vez que diferentes requisitos podem suscitar métodos de aplicação diferentes.

(t1) **Arquiteturas para uma IA de confiança** - conformação dos requisitos da IA de confiança em requisitos, i.e., restrições e políticas, incorporados à arquitetura da IA. Essas restrições e políticas devem refletir-

se em módulos específicos para dar lugar a um sistema global fiável e percepcionado como tal;

(t2) **Ética e Estado de direito por design (X-by-design)** - mecanismos que visam garantir a conformidade entre os princípios abstratos que a IA é obrigada a cumprir e as decisões de aplicação específicas. Deve incluir um mecanismo de suspensão, à prova de falha, e permitir que o funcionamento seja retomado após elucidados os motivos;

(t3) **Métodos de explicação (explicabilidade)** - métodos e modelos que tornam o comportamento e as predições da IA compreensível para humanos. São essenciais para a implantação da tecnologia fiável (vide 7.3.1 Explicabilidade);

(t4) **Testes e validação** - abarca todos os componentes da IA - i.e., incluindo dados, modelos, ambientes e o comportamento do sistema em geral - e devem: ser concebidos e executados por um grupo de pessoas o mais diversificado possível; verificar múltiplos critérios para analisar as categorias testadas segundo diferentes perspectivas; e assegurar que os seus resultados e ações sejam coerentes;

(t5) **Indicadores de qualidade de serviço** – são métricas para avaliar os testes e o treino dos algoritmos, bem como parâmetros tradicionais de avaliação de software, i.e., funcionalidade, desempenho, usabilidade, fiabilidade, segurança, manutenibilidade.

Métodos não técnicos - são os métodos que podem assistir os processos de assegurar e zelar por uma IA de confiança. os métodos apresentados são complementares ou alternativos entre si, uma vez que diferentes requisitos podem suscitar métodos de aplicação diferentes.

(n1) **Regulamentação** - toda regulamentação tangente a fiabilidade da IA (vide 4.5 Leis);

(n2) **Códigos de conduta** - orientações, carta de responsabilidade social, indicadores essenciais de desempenho, ou documentos de política interna que contribuam e apoiem os esforços no sentido da IA de confiança (vide 4.1 Ética);

(n3) **Normatização** - normas aplicáveis ao design, fabrico, práticas empresariais, sistemas de acreditação e códigos deontológicos das profissões. Exemplos atuais são as normas da International Organization for Standardization (ISO) ou a série de normas do Institute of Electrical and Electronic Engineer (IEEE) P7000. No futuro poderá ser adequado adotar um eventual rótulo de «IA de confiança», para normas técnicas específicas, i.e., que a IA respeita critérios de segurança, robustez técnica e explicabilidade (vide

4.4 Normas);

(n4) **Certificação** - são o resultado da aplicação de normas concebidas para diferentes domínios de aplicação e técnicas de IA, adequadamente harmonizadas com as normas setoriais e societais dos diferentes contextos (vide 4.3 Reputação);

(n5) **Responsabilização por meio de conselho de governança** - tanto internos como externos, garantem a responsabilização pelas dimensões éticas das decisões associadas ao desenvolvimento, à implantação e à utilização da IA. Pode-se conformar pela nomeação de uma pessoa responsável pelas questões éticas, relativas à IA, ou um conselho ético interno ou externo (vide

4.4 Normas);

(n6) **Fomento da mentalidade ética** - promover a participação esclarecida do público através da educação básica sobre a IA (vide 4.1 Ética);

(n7) **Participação das partes interessadas e diálogo social** - debater ativa e abertamente, com o envolvimento das partes interessadas, a

participação e o diálogo sobre a utilização e o impacto da IA. Isso contribui para a avaliação dos resultados e das abordagens e pode ser particularmente útil em casos complexos;

(n8) **Diversidade e design inclusivo** - as equipes por trás da IA devem refletir a diversidade sociocultural na qual a IA será inserida, o que contribui para a consideração de diferentes perspectivas, necessidades e objetivos. As equipes devem ser diversificadas não só em termos de gênero, cultura e idade, bem como em termos de experiências profissionais e conjunto de competências.

O capítulo 3 trata da Avaliação de uma IA de confiança. Apresenta uma lista, concreta e não exaustiva, de avaliação da IA de confiança, destinada a operacionalizar os requisitos enunciados anteriormente. (HIGH-LEVEL EXPERT GROUP ON AI, 2019).

Adverte para termos em mente que essa lista de avaliação nunca será exaustiva e que, manter uma IA de confiança, pressupõe um processo contínuo de identificação e aplicação de requisitos, avaliação de soluções e garantia de melhores resultados ao longo do ciclo de vida da IA, e do envolvimento das partes interessadas neste processo.

4.6.4 Ilustração: IA na política

Se até aqui vimos alguns projetos de lei que visam regulamentar e legislar sobre a IA e seu papel na sociedade, ela também foi utilizada para interferir na política, por meio da manipulação social. Os principais meios manipulativos empregados são as Fake News e as técnicas de *microtargeting*.

A produção de *fake news* consiste em informações factuais serem total ou parcialmente trabalhadas, ou retiradas do contexto, acarretando num embuste ou outra falácia lógica.

Percebemos as consequências nocivas do compartilhamento das fake News. Mas, e quando esse compartilhamento deixa de ser direto – e.g., um amigo compar-

tilhando uma história – e passa a ser direcionado para pessoas com maior afinidade e propensão em acreditar nelas? Quais as consequências, éticas, morais, reputacionais, legais e políticas, que podemos esperar?

Microtargeting, i.e., publicidade política direcionada, é uma estratégia de marketing para individualizar ao máximo os consumidores, e utiliza-se de uma comunicação focada em grupos específicos, pensando em qualidade e não em quantidade. A técnica ajuda os políticos a definirem o seu público de um modo específico e descobrir quem seriam os seus potenciais apoiadores.

Para que se possa utilizar esta técnica, nós devemos conhecer o perfil do consumidor minuciosamente. É necessário montar um banco de dados que incluam dados como: idade, sexo, hobbies, comportamento, CEP etc. Assim teremos informações preciosas sobre o público-alvo e poderemos antecipar possíveis resultados. Note que há a necessidade de coleta, armazenamento e processamento desses dados para que possam ser usáveis para o fim proposto.

A fim de obter suas características mais íntimas e seu perfil psicológico, *web crawlers*[49] são empregados na coleta dos dados, que por sua vez são armazenados em grandes data centers e por fim, indexados e analisados por inteligências artificiais para fins de *profiling,* i.e., criação de perfil, e adjeção em segmentos específicos.

Os algoritmos que revelam suas características mais íntimas e seu perfil psicológico têm uma aplicação poderosa: bombardear você com mensagens e ajustar essas mensagens para torná-las mais interessantes e mais relevantes. Fato é que o *microtargeting* se tornou essencial para todos os políticos, que o usam para mandar mensagens específicas a possíveis eleitores e conseguir mais votos.

Além disso, a ferramenta traz outras consequências muitas vezes desprezadas. A redução abissal no custo da comunicação com o eleitor. Desta forma, o diálogo entre candidatos e eleitores se torna muito mais barato podendo ser feito via rede social, e.g., Facebook, Twitter ou WhatsApp.

Como já mencionamos, as mensagens podem ser ajustadas, i.e., torná-las mais interessantes e relevantes para o público. Cada ajuste desse é analisado sobre

---

[49] *Web crawler* é um tipo de robô, i.e., agente de software, que é usado para obter tipos específicos de informações, e.g., minerar endereços de e-mail para spam.

métricas que permitam o candidato tanto fidelizar seus eleitores, arrebanhar novos, e ajustar seu discurso, quanto para evitar determinados debates, temas ou regiões.

Uma possibilidade perniciosa, é a do uso dessa técnica para desencorajar as pessoas a votar, espalhar *fake News*, entre outras coisas. Como descreveremos a seguir, essas técnicas foram empregadas na campanha em prol da saída do Reino Unido da União Europeia.

Brexit é uma abreviação para "*British exit*" ("saída britânica", em tradução literal para o português). Esse é o termo universal para se referir à decisão do Reino Unido de deixar a União Europeia (UE).

Em plebiscito, realizado dia 23 de junho de 2016, eleitores britânicos puderam decidir se o Reino Unido deveria permanecer ou deixar a UE. A maioria, 52%, decidiu que o país deveria deixar o bloco, dando início ao *Brexit* (GUIMÓN e SAHUQUILLO, 2016).

A União Europeia é um grupo formado por 28 países europeus que praticam livre comércio entre si e facilitam o trânsito de sua população para trabalhar e morar em qualquer parte do território. O Reino Unido se tornou parte da EU, na época Comunidade Econômica Europeia, em 1973.

O *Brexit* não aconteceu de imediato, foi inicialmente marcada para o dia 29 de março de 2019 (MORRIS, 2019). Esse prazo não foi cumprido e acabou adiado três vezes, para 31 de janeiro de 2020 (G1, 2019).

Em março de 2017, a decisão de deixar a UE foi notificada ao bloco e o desmembramento se efetivaria dois anos depois (G1, 2019). Março de 2019 chegou e a separação não aconteceu (G1, 2019). Neste período um acordo de retirada foi esboçado pela então primeira-ministra britânica Theresa May e foi rejeitado três vezes no Parlamento do Reino Unido. Isso a levou a deixar o cargo em junho.

Em seguida, Boris Johnson foi eleito chanceler (BBC NEWS, 2019). Durante sua campanha, Johnson prometeu que o Reino Unido sairia da UE dentro do novo prazo (31 de outubro de 2019) (RFI, 2019), com ou sem acordo. No início de setembro de 2019, o Parlamento britânico aprovou uma lei que, na prática, impedia um *Brexit* sem acordo (DW, 2019).

Em outubro Johnson solicitou à UE um novo prazo para o divórcio (31 de janeiro de 2020) (AFP, 2019). Sem consenso entre parlamentares sobre o *Brexit*, Johnson convocou para dezembro novas eleições gerais. Como resultado, o Partido Conservador obteve a maioria das cadeiras no parlamento. Tendo o Parlamento a seu favor, Johnson aprovou seu acordo de retirada (RTP, 2019).

Mas, o que teria a decisão dos eleitores britânicos pelo *Brexit* a ver com manipulação social, *Fake News* ou técnicas de *microtargeting*? Como veremos a seguir o principal implicado parece ser a Cambridge Analytica (CA).

Cambridge Analytica foi uma empresa britânica de *datamining*, criada em 2013, com foco em processos eleitorais. Tornou-se conhecida como a empresa que trabalhou para a eleição presidencial de Donald Trump, (BLOOMBERG, 2018) e a favor do *Brexit* (HERN, ALEX;, 2019), visando a saída do Reino Unido da União Europeia.

A empresa esteve no centro de um escândalo mundial sobre a extração ilegal de informações digitais, i.e., sem consentimento, de milhões de usuários do Facebook e como esses dados foram potencialmente usados para *microtargeting* (GRANVILLE, 2018).

A Cambridge Analytica prestou serviços para a campanha Leave.EU e para o United Kingdom Independence Party (UKIP) antes do referendo de 2016, i.e., *Brexit* (DN/LUSA, 2018). A Leave.EU usou dados criados pela Cambridge Analytica para ações on-line de *microtargeting* para influenciar a opinião pública a favor do *Brexit* no referendo de 2016 (HERN, 2019).

Se isso é um ato de interferência (KIRKPATRICK, 2017) ou simplesmente o resultado da polarização do atual cenário político (OSWALD, 2019), o resultado é uma perda de confiança na mídia, em nossas instituições e, eventualmente, em nosso processo democrático.

Fato é que as mídias sociais cresceram em importância e agora podem amplificar a voz de um indivíduo a um nível nunca visto na história da humanidade. Mas, no mundo pós-verdade, como reagem os legisladores quando veem as novas mídias sendo usadas para espalhar *fake News*?

Magistrados e legisladores pedem uma maior regulação sobre a atuação desses grupos especializados em campanhas políticas digitais. A suspeita é que haja pouca transparência sobre como as métricas (i.e., os "*likes*" do Facebook, Twitter e Google) e demais dados são obtidos, usados e manipulados. Alegam também uma falta de controle sobre o funcionamento dessas campanhas políticas digitais.

O Governo Federal promulgou a Lei 13.834/2019 que torna crime a denunciação caluniosa com finalidade eleitoral. O texto prevê pena de prisão de dois a oito anos, além de multa, para quem acusar falsamente um candidato com o objetivo de afetar a sua candidatura, versando seu art. 2º parágrafo 1 que, – a pena é aumentada de sexta parte, se o agente se serve do anonimato ou de nome suposto (BRASIL, 2019).

4.6.5 Ilustração: IA na gestão pública

Até aqui vimos os possíveis resultados danosos do uso da IA na interferência política. Mas e quando empregamos a IA na gestão pública?

No primeiro episódio da terceira temporada de Black Mirror, intitulado Nosedive, a protagonista vive em um mundo onde todas as interações sociais e comerciais são pontuadas. Tudo depende da pontuação social (i.e., bens, serviços e até interações sociais diversas) e todos almejam por elevar suas respectivas pontuações. Contudo, o sistema de pontuação sofre de um grande problema: subir de nível é incrivelmente difícil, enquanto descer é rápido e fácil (Black Mirror; Season 3 Episode 1 - Nosedive, 2016).

Apesar de parecer saído desse roteiro, o sistema de crédito social chinês, anunciado em 2014, funciona de forma análoga. Assim como no roteiro da série, a pontuação social de uma pessoa pode subir e descer, dependendo do seu comportamento. Ao contrário da série, no sistema chinês, apenas o governo atribui ou remove pontuação. A metodologia exata nunca foi divulgada, contudo, exemplos de infrações incluem dirigir mal, fumar em zonas proibidas, gastar com supérfluos e publicar notícias falsas ou contra o governo. As punições vão do banimento de voos e

trens (GOH, BRENDA, 2018) ao ingresso na lista negra do governo por não prestação do serviço militar (倪, 雪莹;, 2018).

Inúmeros outros exemplos de políticas públicas envolvendo segurança, saúde, gestão pública etc., poderiam ser apresentados aqui para convencer o leitor do impacto da inteligência artificial na política. Como veremos a seguir, a comunicação, a educação e a formação desempenham um papel importante na difusão dessas políticas, leis e valores e do papel social de cada um nessa nova realidade permeada pela IA, além de expectativas realistas sobre ela.

4.7 EDUCAÇÃO

A comunicação, a educação e a formação desempenham um papel importante, tanto para assegurar uma ampla difusão dos conhecimentos sobre o potencial impacto da IA como para sensibilizar as pessoas para o facto de que podem influenciar o desenvolvimento societal.

À medida que a inteligência artificial evolui, crescem as controvérsias causadas pelo seu uso. De armas autônomas letais (LAWs) ao Netflix, a IA está presente em uma gama de produtos e serviços, e com ela, novos dilemas éticos para os quais não há respostas fáceis surgem nos diversos setores em que é empregada.

Portanto, é crucial que os profissionais, e aspirantes, tenham uma educação imbuída das responsabilidades que acompanham o uso da IA. Abaixo, abordaremos alguns dos principais cursos oferecidos por instituições de ensino, como Instituto de Tecnologia de Massachusetts (MIT) e a Universidade de Stanford, e instituições, como Microsoft e a IEEE, em seus esforços para difundir boas práticas no desenvolvimento ético da IA.

O primeiro grupo de cursos visa atender o grande público, i.e., não requerem conhecimento prévio sobre a área. Apresentam conceitos prementes da área como imparcialidade, *accountability* e transparência e abordam o assunto de forma holística.

O curso de ciências e dados e ética oferecido pela Universidade de Michigan aborda o impacto lato sensu da ciência de dados na sociedade moderna sobre os princípios da imparcialidade, *accountability* e transparência à medida que fomenta a importância de um conjunto compartilhado de valores éticos. O curso atende de cientistas de dados à curiosos que se aventuraram pela área de mineração de dados uma vez que nenhum conhecimento prévio específico é necessário e aborda assuntos como: quem possui os dados; como valorizamos os diferentes aspectos da privacidade; e o que significa ser imparcial.

Ainda nessa seara, a IEEE oferece o curso “Responsible Innovation In The Age Of AI”. Num nível introdutório e voltado para gestores, este curso mostra como, ao implementar a ética aplicada em seus processos, as empresas podem aprender a incorporar holisticamente a IA em seus processos de inovação, obtendo lucros e honrando princípios éticos.

Através de aulas expositivas para nivelamento, estudos de caso, discussão de artigos, debate com convidados e seminários em grupo, cursos como “The Ethics of Artificial Intelligence” da Vanderbilt University, “Ethics of AI: Safeguarding Humanity” e “The Ethics and Governance of Artificial Intelligence” do MIT e “Ética e Inteligência Artificial” da Universidade Federal de Pernambuco visam incutir nos futuros profissionais, valores técnicos e éticos. Os tópicos específicos abordados no curso incluem as implicações técnicas, seguracionais e econômicas da IA. Subáreas específicas incluem transporte, manufatura, jornalismo, assessoria jurídica e aplicações militares; vieses nas IA, que podem reforçar o preconceito e a discriminação; explicabilidade de IA, que visa minimizar a opacidade de alguns modelos de IA; ferramentas de suporte a decisão dotadas de IA em áreas como planejamento ambiental e de recursos; IA e personalidade, abrangendo as implicações teológicas da IA.

Abordagens interdisciplinares também são oferecidas. O curso "Ethics And Law In Analytics and AI" da Microsoft ensina a aplicar estruturas éticas e legais no desenvolvimento de IA. Já o “The Economic Advantage Of Ethical Design” oferecido pela IEEE foca nos ganhos econômicos para a empresa que opta por adotar um design ético para suas IA.

Todos os cursos abordados até aqui visam capacitar alunos, profissionais e empresários nas boas práticas do desenvolvimento ético da IA e seus respectivos proventos legais, econômicos, políticos e sociais. Mas e quanto a educação básica?

Ao se perguntar "O que os alunos [do ensino médio] já sabem sobre inteligência artificial nessa idade?" (MICHELLE MA, 2019) Blakeley H. Payne desenvolveu um curso de IA para alunos do ensino médio com a ajuda de professores e pesquisadores de ciência da computação da Harvard Graduate School of Education. O intuito do curso era desmistificar "o que é IA" e conscientizar as crianças sobre como os sistemas de IA mediam sua vida cotidiana, do YouTube e Alexa da Amazon à pesquisa no Google e mídias sociais.

Testado em outubro de 2018 com cerca de 225 alunos da quinta à oitava série da David E. Williams Middle School, em Coraopolis, Pensilvânia, o currículo de uma semana e meia ensina às crianças conceitos sobre IA, como viés algorítmico, recomendação algorítmica e *deep learning*. Cada lição de 45 minutos normalmente incluiu uma breve palestra e demonstração, seguida de uma atividade em grupo e discussão aberta (PAYNE, 2019). O próximo passo foi construir um vocabulário comum: perguntar "o que é IA" e desmistificá-la, porque a inteligência artificial é antropomorfizada cada vez mais pelas empresas de mídia e tecnologia.

Segundo a autora, a literatura sugere que o ensino médio é uma época em que os alunos começam a desenvolver pensamentos de raciocínio moral mais elevados e complexos. Além disso, é geralmente nessa faixa etária que os alunos têm seu primeiro telefone celular ou sua primeira conta em mídia social (BETSY MORRIS, 2018), sendo importante esse tipo de intervenção nessa fase de maior independência.

Payne sinaliza para a necessidade de o educador realizar uma previa autoeducação, não se abater com a aparente complexidade do assunto e ter conversas honestas com seus alunos sobre quais tecnologias eles estão usando, a fim de evitar um maior descompasso entre os exemplos ensinados e a realidade do aluno.

Se iniciativas educacionais com foco na interação IA-sociedade ainda principiam, outras, com foco no desenvolvimento sobejam. A AI4Children da Dalton Learning Lab oferece serviços que permitem ensinar Inteligência Artificial a crianças

usando a linguagem de programação Scratch. Já a ReadyAI oferece, para pais, educadores e instituições de ensino, uma gama de produtos e serviços voltados para o ensino da programação de IA para crianças e jovens adultos.

E quando a ação educacional visa o grande público? Quer seja para difundir uma série de conceitos fundamentais, quer sejam as últimas novidades na área, diversos são os meios utilizados, que vão da tradicional mídia (e.g., revistas, jornais, programas de tv) até as novas (e.g., Youtube, sites e plataformas de EAD).

A primeira dessas categorias, dirigidas ao grande público, faz uso da mídia especializada de massa e, geralmente, visa difundir produtos e serviços que incorporem IA. Podemos citar, como exemplo de anúncio de serviço, o comercial da BIA (a IA do banco Bradesco) que promete revolucionar seu acesso aos serviços do banco e facilitar sua vida (BRADESCO, 2018). Outro exemplo bem interessante é do comercial da Alexa (assistente pessoal da Amazon), veiculado na final do Super Bowl no ano de 2020 (AMAZON, 2020). É interessante notar que, via de regra, as IA são apresentadas de forma antropomórfica, prestativas e bem-humoradas nesses comerciais. Elas são quase uma antítese das IA imortalizadas no cinema como H.A.L. 9000 (2001: A Space Odyssey, 1968) e Skynet (The Terminator, 1984).

Na categoria de novas mídias, o destaque fica por conta do YouTube. Nele encontramos diversos vídeos avulsos falando sobre o tema. Nele encontramos iniciativas como o CrashCourse AI, uma iniciativa crowdfund (i.e., colaboração coletiva), que visa passar para o público os principais conceitos da IA incluindo algumas práticas de programação, em 20 vídeos. Outro grande filão hospedado no serviço são os canais de ensino de especificação, modelagem e programação de IA.

Com foco na didática através do ensino a distância (EAD), existem mais de 20 portais especializados que oferecem cursos gratuitos, e pagos, aos interessados. As principais EAD são o Coursera, com 541 cursos sobre a temática da IA em seu catálogo, (COURSERA, 2020) e a Udemy, com 1339 (UDEMY, 2020). A maioria desses cursos são gratuitos e ofertados por pessoas com experiencia tácita na área; mas, as principais universidades (e.g., MIT, Stanford, Harvard, Berkley) e empresas (e.g., IBM, Microsoft, Google) também disponibilizam cursos oficiais.

Todos os meios educacionais apresentados até aqui visam, de alguma forma, ensinar sobre IA, e.g., suas características, benesses, funcionamento etc. Quer seja para o grande público ou voltada para uma classe específica (e.g., programadores, legisladores, empresários, economistas, estudantes) cada um deles tem por finalidade moldar ou aprimorar o nosso entendimento sobre uma parte da IA. Os principais ganhos acabam sendo desmistificar e dilucidar o assunto perante o público a que se destina e, transitivamente, aos demais. Esse ganho de formação acaba, mesmo que limitadamente em alguns casos, implicando numa diminuição da ansiedade e temores, e.g., discriminação, desemprego, manipulação, mas podem acarretar novas celeumas.

Ao tomarmos a definição de *accountability* como a efetivação de uma promessa, fica fácil entender por que um indivíduo deve responder e explicar suas ações. Há uma tendência crescente das empresas em se comportarem como bons cidadãos globais. Embora um aspecto disso possa ser uma pretensa liderança ética, outro é a criação de benefícios e a prevenção de danos reputacionais.

Como mencionado ao longo do capítulo, existem alguns aspectos sociais importantes que são impactados pelo desempenho da IA e que a maioria das pessoas concorda serem sempre responsabilidade dos intendentes. No entanto, existem muitos exemplos de situações em que as organizações não aceitaram qualquer responsabilidade ou prestação de contas pelos impactos sociais ligados às operações da IA.

Há um movimento crescente em direção a gastos e investimentos éticos, o que, por sua vez, promove o escrutínio público da prestação de contas das empresas. Por meio dos dados coletados e divulgados propiciamos aos diversos stakeholders, dentro e fora da organização, uma base para desenvolverem suas análises factuais e tomarem eventuais decisões informadas.

As razões por trás da produção dos balanços são multifacetadas. Embora as duas principais considerações sejam os impactos financeiro e reputacional, existe um rol de benefícios internos e externos que podem incluir:

(I1) Maior compreensão dos riscos e oportunidades;

(I2) Diferenciação entre o desempenho financeiro e não financeiro;

(I3) Influenciar a estratégia e a política de gestão de longo prazo e os planos de negócios;

(I4) Agilizar processos, redução de custos e melhoria da eficiência;

(I5) Benchmarking e avaliação do desempenho da IA em relação a leis, normas, códigos, padrões de desempenho e iniciativas voluntárias;

(I6) Evitar ser implicado em falhas divulgadas;

(I7) Comparar o desempenho internamente e entre organizações e setores.

(E1) Mitigar – ou reverter – impactos negativos;

(E2) Melhorar a reputação e a fidelidade à marca;

(E3) Permitir que as partes interessadas externas entendam o verdadeiro valor da organização e os ativos tangíveis e intangíveis;

(E4) Demonstrar como a organização influencia e é influenciada pelas expectativas sobre o desenvolvimento da IA.

Tendo considerado a questão de por que coletar e divulgar informações específicas sobre a IA, a próxima questão que podemos considerar é ' A quem reportar?”

5 QUEM SÃO OS STAKEHOLDERS CONTEMPLADOS POR ESSAS INFORMAÇÕES?

Conforme enfatizado anteriormente, os gerentes precisam fazer julgamentos sobre a quem eles devem prestar contas e se devem incluir todos os stakeholders que podem impactar a organização. Ao longo do tempo, houve um aumento na demanda por informações sociais e técnicas pertinentes a IA por diferentes grupos, particularmente em relação a questões como preconceito algorítmico, déficit de confiança na tecnologia, poder computacional, conhecimento limitado, escassez de dados, segurança, perda de emprego, consequências não intencionais, privacidade e segurança de dados.

Não há muita discordância sobre que tipo de entidade pode ser uma parte interessada. Pessoas, grupos, bairros, organizações, instituições, sociedades e até mesmo o meio ambiente são geralmente considerados como interessados reais ou potenciais. Portanto, o grupo de stakeholders a quem as organizações se reportam ampliou ao longo do tempo.

Stakeholders, ou partes interessadas, são organizações ou grupos sociais de qualquer porte que atuam em vários níveis (doméstico, local, regional, nacional, internacional, privado e público), que têm participação significativa e específica em um determinado conjunto de recursos, podendo afetar ou ser afetados pela gestão desses recursos (FREEMAN, 1984). Em suma, são todos aqueles que precisam ser considerados para alcançar os objetivos da IA, cuja participação e apoio são cruciais para o seu sucesso, e são diretamente, ou indiretamente, impactados por ela.

Antes de começarmos a analisar os stakeholders, suas demandas, prioridades e interrelações, é interessante que retomemos nossa dinâmica de pacto. Começaremos analisando um caso ideal onde a demandada (concedente) e a demandante (solicitante) lidam diretamente entre si.

Nesse cenário, o solicitante precisa garantir que seu pedido seja o mais claro possível para ambas as partes. Deve explicar as ideias e reflexões por trás do pedido. A demanda deve ser expressa em uma negociação e nunca imposta. As opiniões da parte solicitada devem ser requeridas e genuinamente consideradas e deve

haver tempo para que todos reflitam sobre a realização do pacto. A concedente, caso concorde, deve assumir o que foi prometido considerando riscos e mitigações eventuais. A negociação se encerra quando há uma conclusão satisfatória.

A fim de tornar a petição a mais clara possível, o solicitante deve planejar o pedido. Deve-se expressar com clareza o que exatamente se deseja alcançar, i.e., tarefas e metas a serem cumpridas, e como cada uma das partes pode e deve contribuir para o sucesso da empreitada.

Por vezes, durante a negociação, transformar um objetivo de longo prazo em um pedido específico de prazo mais curto se mostra uma alternativa melhor. Um pedido bem formulado por parte do demandante deve responder as seguintes perguntas:

1) O que estou almejando? O que e por que quero que seja feito?
2) Como formular meu pedido de forma clara?
3) Como alcançar meu objetivo. Como dividir em tarefas factíveis?
4) Como saberei se alcancei meu objetivo? Como mensurar o progresso da tarefa?

Ao estabelecer de forma clara o que se espera do demandado, tarefas e objetivos a serem cumpridos e métricas que assegurem a realização do trato, o demandante está apto a proceder para a etapa seguinte, a negociação do pacto. Como solicitante, seu objetivo é estabelecer uma cultura de feitura com a concedente.

O elã de concretização do pacto é fomentado por uma interação de pequenas solicitações que erigirão o sucesso por meio de uma cadeia de êxitos. O demandante, portanto, deve criar uma relação de confiança, conhecimento e propósito compartilhado com todas as concedentes com quem pretende trabalhar.

A motivação de realização é maior quando as pessoas atuam com objetivos auto estabelecidos e compartilhados. A fase de negociação envolve, principalmente, dirimir as distâncias entre o que o solicitante quer e o que a concedente acredita ser possível realizar.

Assim, devemos acrescentar outras duas perguntas normativas ao nosso roteiro de formulação do pedido:

5) Quem é a melhor pessoa para conseguir que isso seja feito para mim?
6) Como lidar com objeções, impossibilidades, renegociações e o que posso ou não abrir mão?

Quando um plácito é feito e não plenamente satisfeito, usualmente se observa uma falta, ou deficiência, de comprometimento por parte da demandada. Objetivos impostos e não alcançados são colocados na conta do solicitante, e.g., sob a justificativa de inexequibilidade ou má formulação do trato.

Tão importante quanto ser claro quanto ao objetivo é não transformar um pedido numa ordem. Falar para alguém o que fazer não gera o mesmo comprometimento que um acordo. O acordo envolve um senso moral e principalmente reputacional.

Uma conversa de comprometimento é uma discussão de duas mãos. O processo de solicitar a opinião da outra parte, transforma o ato em uma conversa de comprometimento. Trabalhar junto à concedente permite que esta assuma o controle sobre o resultado do pacto. Faculta que contribua para as soluções ou mitigações possíveis bem como expressar eventuais preocupações e incertezas sobre sua capacidade de satisfazer o trato.

Enquanto concedente, sempre que houver alguma ambiguidade deve-se inquirir até findá-la. Você deve sempre observar fatores como risco, interdependências, e outros fatores que impactem diretamente a exequibilidade do trato.

Entendemos risco como a probabilidade de insucesso de determinado empreendimento, em função de acontecimento eventual, incerto, cuja ocorrência não depende exclusivamente da vontade dos interessados. Isto é, qualquer incidente que possa vir a ocorrer e pôr tudo a perder.

A prevenção e redução de riscos consistem num dos principais elementos que diferenciam as pessoas entre aquelas que obtêm resultados e aquelas que, por vezes, falham em seus objetivos (MCCLELLAND, 1961). Pessoas que não dominam a avaliação de riscos se classificam em duas categorias. As superotimistas, que possuem uma postura entusiasmada em relação ao risco, superestimam suas capacidades e subestimam os riscos. As pessimistas, hesitam em assumir alguma tarefa além daquelas mais simples e seguras.

Tentar demais sem obter resultados ou se recusar a assumir tarefas um pouco mais exigentes prejudicam uma boa reputação e, portanto, deve-se encontrar um equilíbrio entre esses extremos. Ao dominar a avaliação dos riscos você elimina fatores como sorte, acaso, esperança ou medo e trabalha com cenários previsíveis, i.e., onde você antecipou e mitigou a maior parte do que pode vir a acontecer. Isso o prepara para quando algo completamente inesperado ocorrer, você estará preparado para atuar prontamente e lidar com a questão.

Atentemos para o fato de que um processo detalhado de antecipação e mitigação de riscos perdura durante todo o pacto. Além disso, outros fatores devem ser avaliados, e incorporados ao planejamento de risco, antes de nos comprometermos, e.g., complexidade, cronograma e experiência.

Promessas mais simples são mais fáceis de serem cumpridas e, portanto, preferíveis. Criar uma série de tarefas simples e factíveis é melhor que uma única tarefa complexa. Pactos feitos para serem cumpridos em períodos muito extensos apresentam um risco maior (i.e., sujeitas a mais circunstâncias agravantes) que prazos mais curtos. Outro fator é a experiencia. Se você já desempenhou a tarefa antes, conhece as partes envolvidas e como elas trabalham, será mais fácil cumprir a promessa.

Accountability é um pacto entre duas pessoas contudo, nem sempre podemos reunir demandada (concedente) e demandante (solicitante) para lidarem diretamente entre si. Precisaremos depender de outras pessoas para ajudá-las. No entanto, ao envolvermos outras pessoas no pacto, os riscos se tornam cada vez maiores.

É importante conhecer as capacidades e negociar as prioridades e o comprometimento de cada stakeholder e assim ter uma visão geral do contexto em que o pacto será operado. Idealmente, tente obter o comprometimento das pessoas envolvidas de forma individual em vez de equipes, i.e., peça a pessoa para fazer uma promessa a você também. Isso o colocará na posição de concedente e solicitante ao longo de uma cadeia de promessas, mas ao dar sua palavra, você respaldará seu julgamento de qualquer interdependência, i.e., a responsabilidade não pode ser transferida.

Para uma melhor gestão dos stakeholders, de seus interesses e viabilizar o pleno cumprimento do pacto, precisaremos empregar alguma metodologia de análise das partes interessadas. A análise das partes interessadas identifica todos os stakeholders que têm interesse nas questões relacionadas à IA. O objetivo da análise é desenvolver uma visão estratégica do cenário humano e institucional e das relações entre as diferentes partes interessadas e as questões com as quais eles mais se preocupam.

Mapear seus stakeholders ajuda você a avaliar seus relacionamentos com eles para que você possa organizar estratégias de comunicação e esforços de engajamento adequados. Permite conhecer os interesses de seus stakeholders, posições a favor ou contra uma determinada política, se eles serão beneficiados ou prejudicados por um projeto ou política, alianças e conflitos com outras partes interessadas e o grau de envolvimento no processo.

Em outros termos, esquematizar o ajudará a entender que tipo de stakeholder um indivíduo ou organização pode se tornar e a influência e importância dele para o sucesso do pacto. Por isso, é importante realizar a análise antes e durante o ciclo de vida da IA, i.e., de forma contínua, para facilitar a construção de alianças e prever e prevenir possíveis conflitos.

A análise dos stakeholder pode ser definida como categorizar e implantar ações relacionadas às partes interessadas com base em um sistema organizacional interno (níveis de engajamento, prioridade, áreas problemáticas ou função de trabalho), i.e., de acordo com o que faz sentido para sua organização. Existem várias maneiras de realizar essa análise, e.g., workshops, grupos focais e entrevistas são três abordagens comuns, e você pode combiná-las de acordo com as necessidades e evolução do pacto.

No geral, podemos dividir o mapeamento das partes interessadas em quatro etapas diferentes: identificação, categorização, priorização e comunicação.

Identificação – nessa etapa, você deve identificar todas as partes interessadas em potencial (pessoas, grupos, organizações e instituições afetadas pela IA), i.e., aqueles que têm influência sobre ela ou têm interesse ou preocupa-

ção em seu sucesso. Seja o mais granular possível ao arrolar todas. Abordaremos como em 5.1 Identificação dos stakeholders.

Categorização – nessa etapa você deve agrupar os stakeholders listados na fase de identificação nas classificações pertinentes ao tema de categorização. Por exemplo, se o tema for poder vs. influência, os stakeholders serão categorizados numa dessas quatro categorias: baixo nível de interesse e baixo poder; baixo nível de interesse e alto poder; alto nível de interesse e baixo poder e alto nível de interesse e alto poder. Durante o mapeamento pode ocorrer a descoberta, ou necessidade, de um stakeholder não listado, que deverá ser acrescido a lista. Outro problema que pode surgir é uma concentração muito grande dos stakeholder em poucas categorias, o que pode denotar uma deficiência no sistema de classificação. Estudaremos sobre em 5.2 Categorização dos stakeholders.

Priorização – nessa etapa você deve priorizar as principais partes interessadas e começar a tratativa com elas, se possível, no início do projeto (vide 5.3 Priorização dos stakeholders). Existem diferentes maneiras de priorizar as partes interessadas. Por exemplo, os stakeholders classificados como alto nível de interesse e alto poder, e.g., clientes e consumidores, possuem um imenso poder sobre as decisões e os resultados sendo vital portanto, uma comunicação. Abordaremos as principais maneiras em 5.4 Técnicas de mapeamento de partes interessadas.

Comunicação – nessa etapa, é elaborado um plano de comunicação para engajar todas as principais partes interessadas. Não existe uma receita única que possa atender a todas as situações possíveis, mas os bons princípios informacionais já aventados devem nortear a confecção do plano. Outro ponto que deve ser salientado é que o plano de comunicação deve contemplar todos os níveis, i.e., do baixo nível de interesse e poder até o alto nível de interesse e poder. Citaremos como em 5.5 Comunicação com os stakeholders.

5.1 IDENTIFICAÇÃO DOS STAKEHOLDERS

A identificação inicial das partes interessadas consiste em listar grupos conhecidos por influenciarem ou serem impactados pela IA. Esse processo fornecerá uma base importante para expandir a quantidade de stakeholders conhecidos, bem como começar a analisar os relacionados.

Para analisar os grupos de stakeholders, você pode começar estudando alguns cenários específicos de ameaça e oportunidade e elicitar as principais partes interessadas associadas a cada um deles. Alternativamente, você pode arrolar todos os atores com potencial interesse na IA, sem limitar a lista com base no fato de você saber a priori que o grupo terá interesse ou não, e vinculá-los a fatores específicos de ameaça e oportunidade. Mais tarde, durante a análise e engajamento das partes, você terá a chance de confirmar se os grupos têm uma participação relevante ou não.

A listagem inicial pode ser realizada por meio de brainstorming, grupos focais ou entrevistas e guiados por um roteiro de análise de stakeholders. A seguir, algumas perguntas que podem conter um roteiro:

- Quem será afetado pelas atividades da IA?
- Quem poderá influenciar os resultados da IA?
- Quem são os potenciais apoiadores e quem seriam os opositores ou desinteressados?
- Que parcerias podem ser construídas em torno das questões envolvidas?
- Quem será responsável por gerenciar os resultados da IA?
- Quais serão os usuários da IA?
- Quem pode facilitar ou impedir os resultados por meio de sua participação ou não?
- Quem são as partes vulneráveis?
- Como estão sendo ameaçadas as partes vulneráveis?
- ‘Quem é mais dependente dos recursos em jogo?’, ‘Isso é uma questão de subsistência ou vantagem econômica?’ e ‘Esses recursos são substituíveis por outros?’

- Quem possui reivindicações – incluindo jurisdição legal e uso costumeiro – sobre os recursos em jogo? Vários setores do governo e departamentos ministeriais estão envolvidos? Existem órgãos nacionais e/ou internacionais envolvidos por causa de leis ou tratados específicos?
- ‘Quem são as pessoas ou grupos mais bem informados e capazes de lidar com os recursos em jogo?’, ‘Quem está gerenciando esses meios?’ e ‘Com que resultados?’
- Existem grandes eventos ou tendências atualmente afetando as partes interessadas?
- Houve alguma iniciativa semelhante? Em caso afirmativo, até que ponto foi bem-sucedido? Quem estava no comando e como as partes interessadas responderam?

As perguntas acima visam exclusivamente ilustrar alguns dos tópicos a serem abordados numa dinâmica de análise de stakeholders. Elas devem ser adaptadas e acrescidas conforme a necessidade e especificidades do projeto a ser desenvolvido.

5.2 CATEGORIZAÇÃO DOS STAKEHOLDERS

Após a elicitação dos stakeholders, inicia-se a fase de tratamento dos dados. Para tal, começaremos por classificar os stakeholders de acordo com certos critérios, e.g., proatividade, escopo, influência, relevância etc. É interessante que adotemos alguma forma de arrolar essas informações que nos permita uma fácil localização, comparação e manutenção dos dados. O exemplo abaixo é meramente ilustrativo, mas pode servir como base para uma futura adaptação.

Tabela 2 – Exemplo de matiz de análise de stakeholders

Grupo	Stakeholder	Participação	Potencial	Vulnerabilidades	Prioridade

Fonte: Johnson & Scholes (1993)

Os stakeholders podem ser ativos ou passivos segundo a proatividade. Além disso, podemos ter partes interessadas internas e externas. As partes interessadas internas são pessoas que estão participando da construção da sua IA. Seu nível de envolvimento pode variar, mas todos têm influência porque fazem parte da organização. Geralmente compõem o rol de partes interessadas internas: executivos, gestores, projetistas, desenvolvedores, testadores etc. Já as partes interessadas externas são aquelas que serão, primordialmente, impactadas pela IA, embora não participem diretamente da criação da IA. Geralmente compõem o rol de partes interessadas externas: usuários, legisladores, acionistas, parceiros etc.

Essas partes podem apresentar diferentes níveis de envolvimento. Portanto, podemos dividi-las em três categorias diferentes de stakeholders: primários, secundários e terciários.

Stakeholders primários – são afetados positiva ou negativamente pela IA. Essas são as pessoas que têm a conexão mais forte com as atividades e os resultados da IA.

Stakeholders secundários – pessoas que não são diretamente afetadas pela IA. No entanto, eles podem ser afetados indiretamente.

Stakeholders terciários – aqueles que são minimamente afetados pela IA.

Além destas classificações, as partes interessadas podem ser agrupadas com base em outras categorias pertinentes à finalidade, e.g., departamentos governamentais, políticos, formuladores de políticas/consultores (local nacional, internacional), aqueles envolvidos com estratégias nacionais/regionais relevantes; organizações não governamentais; negócios e indústria; grupos profissionais; educadores; meios de comunicação; público em geral; usuários etc. A primeira coluna deve ser preenchida com o grupo ao qual o stakeholder (coluna 2) pertence.

É claro que um stakeholder pode pertencer a mais de um grupo. Contudo, é preferível que seja alocado em apenas um, o mais relevante, e os demais sejam tratados na coluna 4, potencial. Por exemplo, um dos programadores da IA pode ser usuário dela e membro de uma organização não governamental afetada por ela. Nesse caso, é preferível agrupá-lo pelo cargo, dada a relevância dele para o ciclo de

vida da IA, e explicitar as demais categorias na 4ª coluna, e.g., usuário e potencial representante da ONG.

Para preencher a segunda coluna da matriz, stakeholder, liste as partes interessadas arroladas nas perguntas. Em seguida, descreva a participação de cada parte na terceira coluna.

A participação refere-se à natureza e limites da atuação de cada parte interessada (e.g., meios de subsistência, lucro, estilos de vida, valores etc.), e a base dessa participação (e.g., propriedade, responsabilidades administrativas ou legais, direitos, obrigação social etc.). Em seguida, para cada parte interessada, descreva seu papel potencial no projeto na coluna 4.

Anote na coluna 5 se o stakeholder é vulnerável. As partes interessadas vulneráveis[50] carecem do reconhecimento ou da capacidade de participação dos esforços colaborativos em igualdade de condições, e uma atenção especial deve ser destinada para garantir e possibilitar sua participação. Certifique-se de que os stakeholders vulneráveis, que normalmente não têm voz nas discussões e decisões políticas, sejam incluídos na divulgação.

Na última coluna, registre quem são os principais stakeholders. Ou seja, aqueles que por reivindicação, direito, poder, autoridade ou responsabilidade são centrais para a iniciativa em questão. A participação deles é fundamental.

Por último, é interessante atribuir algum código para se referir a cada parte interessada de maneira inequívoca. Isso permite não só evitar problemas de comunicação e retrabalho, como nos permitirá estabelecer toda uma linhagem inequívoca entre solicitantes e demandados.

Não obstante, nossa meta é conceber uma visão estratégica do cenário humano e institucional impactado pela IA e deslindar as relações entre as diferentes partes interessadas. Dessa forma, precisamos mapear não só essas relações, mas a qualidade, e.g., influência e relevância, delas.

---

[50] Estes podem incluir grupos minoritários, mulheres, jovens, grupos indígenas, LGBTQIA+, minorias étnicas e outros grupos empobrecidos ou desprivilegiados.

5.3 PRIORIZAÇÃO DOS STAKEHOLDERS

Há muitas maneiras de priorizar seus stakeholders. É importante ranquear as partes interessadas de acordo com o que você está tentando realizar em sua estratégia de engajamento. Ao decidir como ordenar seus stakeholders, considere estas questões:

- Quem é diretamente responsável pelas decisões sobre questões importantes para o projeto da IA?
- Quem ocupa cargos de responsabilidade nas organizações interessadas?
- Quem é influente na área (áreas temáticas e correlatas)?
- Quem será afetado pela IA?
- Quem irá promover/apoiar a IA, desde que esteja envolvido?
- Quem obstruirá/impedirá a IA se não estiver envolvido?
- Quem esteve envolvido no passado?
- Quem não esteve envolvido até agora, mas deveria estar?

Podemos identificar os diferentes stakeholders com base em alguns critérios pré-determinados. Podemos fazer isso pelo método ‘Influência’, ‘Probabilidade’ e ‘Meios’ (IPM). Por influência, queremos medir qual a interferência da parte interessada nas decisões e na estratégia da IA. Já probabilidade, se refere a chance de a parte interessada impor suas expectativas, enquanto meios trata sobre os recursos disponíveis para os stakeholders imporem suas expectativas.

A influência das partes interessadas em um projeto pode ser baseada em sua contribuição ou no resultado. Por exemplo, um programador terá um impacto direto como contribuinte. Por outro lado, os usuários serão impactados pelos resultados da IA. No entanto, esses dois stakeholders são importantes quando consideramos o mapeamento das partes interessadas. Em casos extremos, e.g., Microsoft TAI, a influência negativa de uma parte interessada (no nosso exemplo os usuários) levou ao término do projeto com consequências nefastas.

Em segundo lugar, devemos considerar a probabilidade desses stakeholders serem impactados pela IA. Por exemplo, o vídeo "HP computers are racist" (WZAMEN01, 2009) sugere que o software de reconhecimento facial da HP Media

Smart não detecta rostos negros. Quando Wanda, uma mulher branca, fica diante da tela, a câmera faz zoom no rosto dela e a acompanha a medida em que ela se movimenta. Mas quando Desi, um homem negro, faz a mesma coisa, a câmera não reage da mesma forma - não acompanha os movimentos dele. Mesmo que esse tipo de situação seja raro, quando ocorrem, devemos estar preparados e termos as partes afetadas em mente. Isso é útil na execução de ações corretivas.

Por fim, tratamos sobre os meios. Alguns dos stakeholders são mais fortes, enquanto outros são mais fracos. Aqui, definimos a força do stakeholder inteiramente em termos de seu poder ou influência. É interessante, contudo, notarmos o quão ampliado pode ser o poder de uma parte quando sua reivindicação viraliza, como no caso de Desi.

Outro conjunto de critérios que podemos usar para identificar os diferentes stakeholders é a ‘Influência’, ‘Previsibilidade’, ‘Interesse’ e ‘Atitude’ (IPIA). Por influência, queremos medir qual a interferência da parte interessada nas decisões e na estratégia da IA. Previsibilidade, se refere a antecipação das ações das partes interessadas. Já interesse, versa sobre o nível de interesse dos stakeholders enquanto atitude trata sobre o suporte depreendido pelas partes.

Quer você opte por algum dos modelos supracitados, quer você utilize uma coletânea própria de parâmetros, estes devem suportar seu método de mapeamento de partes interessadas eleito. A seguir, veremos algumas das principais técnicas de mapeamento de partes interessadas.

5.4 TÉCNICAS DE MAPEAMENTO DE PARTES INTERESSADAS

As técnicas de mapeamento de partes interessadas que apresentaremos a seguir visam, principalmente, ranquear seus stakeholders da melhor, i.e., mais relevante, maneira possível. Baseando-se apenas num restrito conjunto de atributos, esses métodos podem ser resumidos de maneira simples e serão apresentados de acordo com os critérios abordados. Os critérios usados para mapear as partes interessadas pelos modelos são:

- Poder de influência
- Previsibilidade
- Nível de Interesse
- Atitude

Poder de influência – é a capacidade, ou potencial, da parte interessada influenciar o funcionamento ou os resultados de um projeto. Esse poder pode ser derivado de sua posição de autoridade, recursos, popularidade, dependência. Portanto, uma parte em um relacionamento tem poder, na medida em que tem ou pode obter acesso a meios coercitivos, utilitários ou normativos, para impor sua vontade no relacionamento.

Previsibilidade – é outro critério importante para mapear as partes interessadas. Os problemas podem ser classificados em previsíveis ou imprevisíveis. Da mesma forma, as partes interessadas também podem ser mais ou menos previsíveis em termos de suas ações.

Nível de interesse – uma parte interessada que tem um maior nível de interesse no projeto será mais ativa em seu envolvimento. Por outro lado, os stakeholders com menor interesse no projeto terão um papel mais passivo.

Atitude – dentro do nível de interesse, também podemos categorizar a atitude como positiva ou negativa. As partes interessadas com uma perspectiva positiva apoiarão o projeto, enquanto as que têm uma perspectiva negativa em relação ao projeto podem tentar resistir a ele.

Com tempo, energia e outros recursos limitados para rastrear o comportamento dos stakeholders e gerenciar relacionamentos, os gerentes podem não fazer nada sobre as partes interessadas que possuam baixa pontuação nos atributos de identificação, a gerência pode até mesmo não reconhecer a existência desses stakeholders. Da mesma forma, essas partes interessadas provavelmente não darão atenção ou reconhecimento à empresa.

Por outro lado, quando um stakeholder possui alta pontuação nos critérios avaliados, a reivindicação de tal parte interessada é premente. Os gerentes têm o dever imediato e inequívoco de atender e dar prioridade à petição dessa parte.

Antes de prosseguirmos, no entanto, devemos ter em mente que os atributos das partes interessadas são variáveis, e não estacionários. Esses atributos são uma realidade socialmente construída, onde a consciência e o exercício voluntário podem ou não estar presentes.

5.4.1 Matriz de Poder vs. Interesse

A análise das partes interessadas evoluiu nos últimos anos como uma técnica para analisar os prováveis interesses e ações dos stakeholders (JOHNSON e SCHOLES, 1993). Avaliar a importância das expectativas das partes interessadas é fundamental em qualquer análise estratégica de projeto e consiste em fazer julgamentos sobre três questões (NEWCOMBE, 2003).

• Qual a ***probabilidade*** de cada grupo de stakeholder impor suas expectativas no projeto?

• Se esses grupos têm ***meios*** para fazê-lo. Isso está relacionado com o poder dos grupos de partes interessadas.

• O provável ***impacto*** das expectativas dos stakeholder nas estratégias futuras do projeto.

Para avaliar essas três contingências, Robert Newcomb propôs que as partes interessadas devam ser analisadas segundo dois métodos de mapeamento de stakeholders: a matriz de poder vs. interesse e a matriz de poder vs. previsibilidade. A Matriz de poder vs. interesse é um dos métodos mais comuns de mapeamento de stakeholders. Como você pode ver na tabela abaixo, ao mapearmos duas variáveis nesta matriz, o poder das partes interessadas versus o seu nível de interesse, obtemos quatro grupos com características distintas que abordaremos em seguida.

Tabela 3 – Exemplo de Matriz Poder vs. Interesse

Poder	Nível de Interesse: Baixo	Nível de Interesse: Alto
Baixo	A Esforço mínimo	B Manter informado
Alto	C Manter satisfeito	D Importante

Fonte: Newcombe (2003)

Baixo Interesse e Baixo Poder – quadrante superior esquerdo (**A**), temos o quadrante de baixo nível de interesse e baixo poder. O tipo de stakeholder que se enquadra neste quadrante exigirá um esforço mínimo da empresa. Um exemplo dessa parte interessada seria o público em geral ou um espectador. Essas pessoas teriam um poder mínimo e não teriam interesse ativo em agir a favor ou contra.

Alto Interesse e Baixo Poder – quadrante superior direito (**B**). Nesse quadrante, podemos observar que o nível de interesse é alto, enquanto o poder é baixo. Um dos exemplos desse tipo de stakeholders seriam os trabalhadores contratados. Esses funcionários teriam muito interesse no projeto, mas não têm muita influência sobre as decisões do projeto. Todos esses tipos de partes interessadas não têm muito poder, mas precisam ser mantidos informados.

Baixo Interesse e Alto Poder – quadrante inferior esquerdo (**C**). Nesse quadrante, você pode ver que o poder é alto, mas o nível de interesse é baixo. Um dos exemplos importantes desses tipos de partes interessadas seria o governo. Os governos geralmente não tentam interferir nos negócios do dia a dia das empresas. Porém, quando julgam necessário, possuem poder suficiente para sobrepor-se à administração, investidores etc. É de extrema importância manter esses stakeholders satisfeitos.

Alto Interesse e Alto Poder – quadrante inferior direito (**D**). Neste quadrante, vemos que o poder é alto. Ao mesmo tempo, o nível de interesse também é alto. Esse tipo de situação é muito difícil de lidar se sair do controle. Essas são partes interessadas importantes para nós. É por isso que eles são chamados de 'Importante'. Um dos exemplos mais simples disso pode ser consumidores ou clientes que geralmente possuem imenso poder sobre as decisões e os resultados do projeto.

5.4.2 Matriz de Poder vs. Previsibilidade

Agora vamos analisar a segunda das técnicas de mapeamento de stakeholders proposta por Newcomb. Embora esta técnica não seja tão popular quanto a anterior, é bastante útil pois pode prever o nível de passividade (ativo ou passivo) da parte interessada. No entanto, a previsibilidade pode antecipar certos resultados por parte dos stakeholders.

Tabela 4 – Exemplo Matriz Poder vs. Previsibilidade

Poder	Previsibilidade: Alto	Previsibilidade: Baixo
Baixo	A Poucos Problemas	B Imprevisível Gerenciável
Alto	C Poderoso Previsível	D Perigoso Oportunidade

Fonte: Newcombe (2003)

Alta Previsibilidade e Baixo Poder – quadrante superior esquerdo (**A**). Neste quadrante, vemos que a previsibilidade é alta enquanto o poder é baixo.

Esse tipo de situação não pode ser muito difícil para os gerentes lidarem. Embora qualquer problema oriundo desses stakeholders possa ser problemático, eles podem ser resolvidos com bastante facilidade. Por exemplo, o software de reconhecimento facial da HP Media Smart que não detectava rostos negros poderia ser facilmente evitado ao se considerar a diversidade de usuários do sistema. O treinamento da IA pode demandar um pouco mais de tempo, mas o resultado é um sistema mais robusto e menos propenso a vieses.

Baixa Previsibilidade e Baixo Poder – quadrante superior direito (**B**). Neste quadrante o poder é baixo, mas a previsibilidade também é baixa. Esses tipos de problemas geralmente não podem ser previstos pelos gerentes. Surpresas acontecem, no entanto, estas não ameaçam as operações do projeto. Podemos ter pessoas ou entidades que discordam de certos posicionamentos do projeto ou da empresa e elas podem expressar suas preocupações de diversas maneiras. Esses tipos de contendas são totalmente imprevisíveis. No entanto, elas podem ser tratadas com uma simples resposta ou, às vezes, o problema pode demandar uma ação de relações públicas.

Alta Previsibilidade e Alto Poder – quadrante inferior esquerdo (**C**). Neste quadrante, vemos que a previsibilidade é alta. A poder também é alto. Cada tipo de stakeholder apresenta desafios únicos aos gestores. No entanto, não é muito difícil lidar com eles, pois esses problemas são previsíveis. Podemos tomar o exemplo das regulamentações governamentais. Atualmente, muitos países têm regulamentações muito rígidas sobre tratamento de dados pessoais. No entanto, esses regulamentos são bastante previsíveis e são escritos explicitamente.

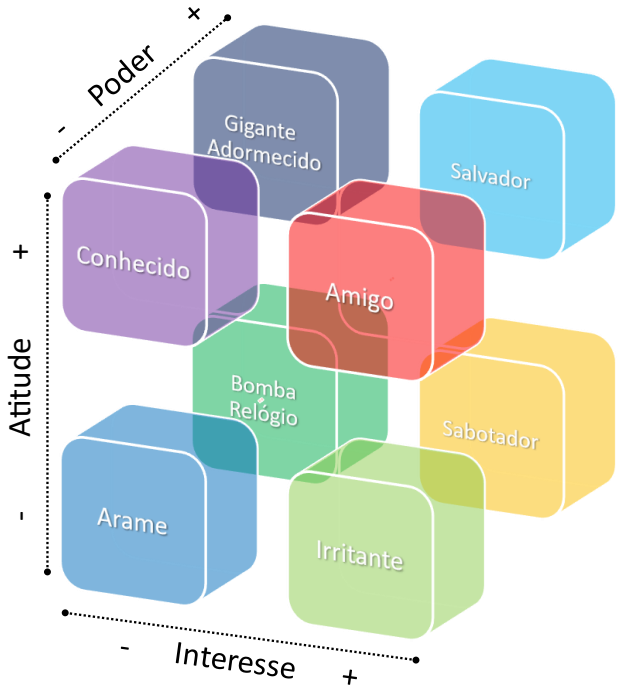
Baixa Previsibilidade e Alto Poder – quadrante inferior direito (**D**). Aqui vemos que o poder é alto, mas a previsibilidade é baixa. Este quadrante apresenta os problemas mais desafiadores para os gestores. Ao mesmo tempo, esses problemas também podem ser encarados como oportunidades. É extremamente importante satisfazer essas partes interessadas.

5.4.3 Matriz Poder vs. Interesse vs. Atitude

Até aqui, as considerações iniciais sobre partes interessadas geralmente são feitas usando uma técnica baseada em duas dimensões, onde dois eixos são rotulados com características de status ou comportamento das partes interessadas e a área entre os eixos (a grade bidimensional) preenchida com os nomes de cada indivíduo ou grupo. Murray-Webster e Simon desenvolveram uma técnica que combina diferentes matrizes numa única e fornece uma visão geral das partes interessadas. Nesse modelo, três dimensões diferentes (poder, interesse e atitude) são analisadas concomitantemente, informando o gerente de uma maneira verdadeiramente significativa (MURRAY-WEBSTER e SIMON, 2006).

Nesta técnica, criamos uma matriz 3D com dimensões 2x2x2. Cada dimensão acomoda uma métrica (poder, interesse e atitude) e suas intensidades (-, +). Como você pode ver no diagrama abaixo, temos oito quadrantes diferentes aqui. As partes interessadas podem ser acomodadas em uma dessas oito caixas.

Tabela 5 – Exemplo Matriz Interesse vs. Atitude vs. Poder



Fonte: Murray-Webster & Simon (2006)

Salvador – poderoso, alto interesse, atitude positiva ou alternativamente influente, ativo, patrocinador. Eles precisam de atenção; você deve fazer o que for necessário para mantê-los ao seu lado - atender às suas necessidades.

Amigo – baixo poder, alto interesse, atitude positiva ou, alternativamente, insignificante, ativo, patrocinador. Eles devem ser usados como confidentes ou caixa de ressonância.

Sabotador – poderoso, alto interesse, atitude negativa ou alternativamente influente, ativo, bloqueador. Eles precisam ser engajados para se desengajar.

Irritante – baixo poder, alto interesse, atitude negativa ou, alternativamente, insignificante, ativo, bloqueador. Eles precisam ser engajados para que parem de 'atrapalhar' e depois sejam 'colocados de volta em sua caixa'.

Gigante Adormecido - poderoso, baixo interesse, atitude positiva ou alternativamente influente, passivo, apoiador. Eles precisam ser engajados para despertá-los.

Conhecido – baixo poder, baixo interesse, atitude positiva ou alternativamente insignificante, passivo, apoiador. Eles precisam ser mantidos informados e comunicados com base em 'transmitir apenas'.

Bomba-relógio - poderoso, baixo interesse, atitude negativa ou, alternativamente, influente, passivo, bloqueador. Eles precisam ser compreendidos para que possam ser “desarmados antes que a bomba exploda”.

Arame – baixa potência, baixo interesse, atitude negativa ou alternativamente insignificante, passivo, bloqueador. Eles precisam ser entendidos para que você possa "observar o seu passo" e evitar "tropeçar".

5.4.4 Mapeamento das partes interessadas por meio do Diagrama de Relevância.

Mitchell, Agle e Wood propuseram que as partes interessadas devam ser analisadas pelas lentes de sua "relevância", i.e., o grau em que os gerentes dão prioridade às reivindicações concorrentes das partes interessadas. Eles propõem que, com base em três atributos de relacionamento, podemos determinar os tipos de stakeholders. Esses atributos são:

(1) o ***poder*** das partes interessadas para influenciar a empresa;

(2) a ***legitimidade*** do relacionamento das partes interessadas com a empresa e;

(3) a ***urgência*** da reivindicação da parte interessada sobre a empresa.

Essa teoria produz uma tipologia abrangente de stakeholders com base na suposição normativa de que poder, legitimidade e urgência definem o campo das partes interessadas: aquelas entidades às quais os gerentes devem prestar atenção.

Com base nessa tipologia, eles propõem ainda uma teoria da relevância das partes interessadas. Nessa teoria, eles sugerem um modelo dinâmico, baseado na tipologia de identificação, que permite o reconhecimento explícito da singularidade situacional e da percepção gerencial para explicar como os gerentes priorizam os relacionamentos com os stakeholders (MITCHELL, AGLE e WOOD, 1997).

Um tema comum entre esta e as demais técnicas anteriores de mapeamento de stakeholders é o 'poder'. Com base na relação de poder, podemos resumir as partes interessadas conforme a Tabela 6.

Ao combinar as diferentes justificativas para a identificação das partes interessadas, os autores vislumbraram a necessidade de avaliar sistematicamente as relações stakeholders-gerente, tanto reais quanto potenciais, em termos da relativa ausência ou presença de todos ou alguns dos atributos: poder, legitimidade e urgência. Concluiu-se, então, que a relevância dos stakeholders será proporcional ao acúmulo de atributos - poder, legitimidade e urgência - percebidos pelos gerentes.

Passamos agora à nossa análise das classes de stakeholders que resultam das várias combinações destes atributos, como se mostra na Figura 2.

Tabela 6 – Classificação de justificativas para a identificação dos stakeholders

Relacionamento Transacional	Transações de troca
	Ações influentes
	Obrigação social ou moral
Predomínio do Stakeholder	A empresa existe por causa dessas pessoas
	A empresa é dependente para sobrevivência, operação ou crescimento
	Indivíduo ou grupo podem ditar termos, influenciar ou dirigir a empresa
Predomínio da Empresa	A empresa é responsável pelo bem-estar do indivíduo ou grupo
	A empresa tem controle ou responsabilidade legal, social ou moral

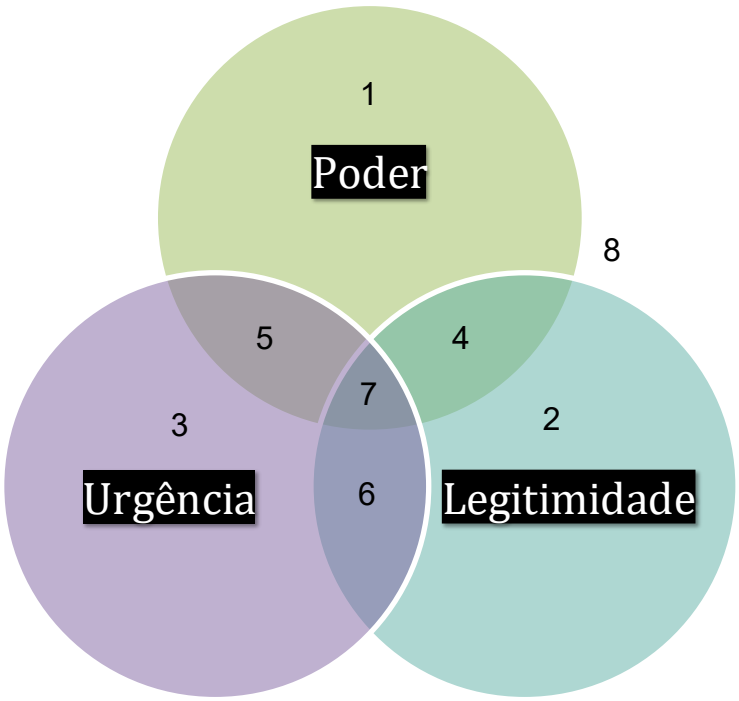
Fonte: Mitchell et al. (1997)

As classes de baixa relevância (áreas 1, 2 e 3), que denominamos partes interessadas “latentes”, são identificadas por desfrutarem de apenas um dos atributos. As partes interessadas moderadamente relevantes (áreas 4, 5 e 6) são identificadas por gozarem de dois dos atributos e, por serem partes interessadas que “esperam alguma coisa”, nós as chamamos de partes interessadas “expectantes”. A combinação de todos os três atributos (incluindo as relações dinâmicas entre elas) é a característica definidora dos stakeholders altamente relevantes (área 7) (MITCHELL, AGLE e WOOD, 1997).

Com base no diagrama de Venn da Figura 2, podemos agora discutir os oito tipos diferentes de partes interessadas. Essas partes interessadas não são organizadas por sua importância, mas por seu mapeamento geral na instituição.

Em suma, uma técnica de mapeamento de partes interessadas deve levar em conta atributos relevantes, não importa quão díspares sejam os resultados. Os gerentes devem conhecer as entidades em seu ecossistema que detêm o poder bem como os afetados que têm a intenção de impor sua vontade.

Figura 2 – Classes Qualitativas de Stakeholders



Fonte: Mitchell et al. (1997)

Qualquer tipo de método de mapeamento só tem valor se puder ser usado de maneira sensata. Estabelecer as posições de indivíduos e grupos na grade é de valor limitado se não for seguido. Claro que o posicionamento pode estar errado com consequentes riscos ao projeto e aos relacionamentos. Mas para qualquer parte interessada que você acredite ser importante, deve-se engajá-la no diálogo, verificar suas suposições e descobrir seus verdadeiros motivadores e preocupações. Esta é a ação que transforma seu mapeamento de stakeholders em análise de stakeholders.

Tabela 7 – Tipologia de Stakeholder

Latentes	1	Adormecidos
	Têm poder, mas não têm legitimidade ou urgência. Isso também significa que esse tipo é passivo.	
	2	Discricionários
	Possuem pouca urgência e poder. No entanto, são conhecidos como partes interessadas discricionárias, pois têm legitimidade.	
	3	Exigentes
	Como o nome sugere, são meticulosas. Esses stakeholders exercem sua influência por causa da urgência. No entanto, eles não têm poder ou legitimidade suficientes para apoiar suas reivindicações.	
Expectantes	4	Dominantes
	São aquelas que têm tanto poder quanto legitimidade. Esses stakeholders exercem certa influência sobre a organização. No entanto, sua influência não é urgente.	
	5	Perigosos
	Têm poder e urgência. Eles podem criar muitos incômodos dentro da organização. Esse tipo precisa ser combatido de forma rápida e correta. O pior é que podem não ter legitimidade ou envolvimento com o projeto.	
	6	Dependentes
	Possuem necessidades urgentes e legítimas. No entanto, eles não têm poder suficiente para exercer essas necessidades. Como resultado, dependem de alguma outra parte interessada para exercer sua influência.	
Definitivos	7	Definitivos
	Detêm todos esses três ingredientes. Em outras palavras, podemos dizer que esses são os stakeholders com o máximo de influência. Como resultado, sua importância na empresa é a maior.	
Potenciais	8	Não Stakeholders ou Stakeholders Potenciais
	Como o nome sugere, não têm voz na organização. Geralmente lhes falta o poder, a urgência e a legitimidade. Esse tipo de pessoa ou grupo pode se tornar parte interessada no futuro caso se qualifique de acordo com os três critérios definidos acima.	

Fonte: Mitchell et al (1997)

5.5 COMUNICAÇÃO COM OS STAKEHOLDERS

Nessa etapa é elaborado um plano de comunicação para engajar todas as principais partes interessadas. Não existe uma receita única que possa atender a todas as situações possíveis, mas os bons princípios informacionais já aventados devem norteara a confecção do plano.

Um plano de comunicação (PC) é uma abordagem holística para fornecer informações com eficiência às partes interessadas. Os planos de comunicação definem quais informações devem ser transmitidas, quem deve recebê-las, quando devem ser entregues, onde (i.e., meio) a comunicação será compartilhada e como essas mensagens serão rastreadas e analisadas. Um bom plano de comunicação deve abordar, no mínimo, os seguintes tópicos:

Metas – O que você quer que sua comunicação alcance?

Conteúdo – Que informação ou apelo à ação conterá esta comunicação?

Cronograma – Quando e com que frequência você entregará esta comunicação?

Canal – Onde você compartilhará esta comunicação?

Métodos – Quais ferramentas/plataformas você usará?

Público – Quem receberá esta comunicação?

Responsável – Quem está encarregado de enviar esta comunicação?

Uma comunicação bem-sucedida deve ser encarada como um processo contínuo, não um evento único. Agora, sabedores de quais informações contemplar nesse roteiro, é importante saber como proceder para elaborá-lo. A seguir, a Tabela 8 – Quinze etapas do processo de planejamento de comunicação lista os 15 passos para criar um plano de comunicação eficaz. Os diferentes campos que aplicam o PC nem sempre usam as mesmas denominações, número de etapas ou o mesmo conjunto de técnicas por etapa, mas todos têm o mesmo objetivo: planejar a comunicação.

Tabela 8 – Quinze etapas do processo de planejamento de comunicação

	Etapa	Descrição
1	Criar uma declaração de missão	Uma descrição curta e escrita do porquê sua empresa existe e atende exclusivamente ao seu mercado.
2	Definir os objetivos de negócios da sua empresa	O que você deseja alcançar. Para cada objetivo listado em seu plano de comunicação, adicione uma nota sobre seu objetivo de comunicação.
3	Identificar os segmentos de público para o seu plano de comunicação	Escolha todos os grupos que se aplicam à sua organização. Use essas informações para escrever um parágrafo para cada segmento de público e adicioná-lo ao seu plano de comunicação.
4	Estabelecer as metas do seu plano de comunicação	O que você quer que sua comunicação realize. Os objetivos de comunicação devem ser tão específicos quanto possível. Dessa forma, você pode mensurá-los e otimizar sua estratégia de comunicação de acordo.
5	Definir sua vantagem competitiva	Fator que o diferencia de seus concorrentes. Você pode vê-lo como “o que você tem que os concorrentes não têm”.
6	Desenvolver mensagens-chave para cada segmento de público	Defina mensagens fundamentais para todos os seus segmentos de público. Eles atuarão como pontos de referência para todas as conversas com pessoas e organizações externas. Escreva um parágrafo de mensagens-chave para cada um de seus públicos.
7	Selecionar canais e frequência para cada segmento de público	‘Onde seu público gosta de passar o tempo, online ou não?’, ‘O que chama a atenção deles?’, ‘Com que frequência eles visitam essas plataformas e lugares?’ e ‘Qual é o máximo que você pode passar sem se comunicar com seu público e ainda permanecer relevante e lembrado?’.
8	Atribuir responsável	Atribua um membro da equipe do projeto em cada canal listado anteriormente.
9	Identificar os principais eventos da empresa e do setor	Crie um calendário com essas datas em sua estratégia de comunicação e facilite a atualização.
10	Criar um plano de distribuição interna	‘Onde seu plano de comunicação viverá diariamente?’ e ‘Como você garantirá que todos na empresa sempre saibam onde ele está e como os atende, incluindo novos contratados?’
11	Criar um plano de treinamento para equipes voltadas para o público	Com base nas lacunas encontradas, construa um plano de treinamento para essas equipes. Certifique-se de incluir o orçamento, especialistas, tópicos e frequência de treinamento.
12	Desenvolver um plano de comunicação emergencial	‘Que tipos de crises são prováveis para o seu negócio?’, ‘Quem é responsável pela comunicação de crise para cada um desses cenários?’, ‘Que informação você deve transmitir agora ou mais tarde?’, ‘Quais cenários exigirão aconselhamento jurídico?’, ‘Quem é o

		porta-voz de cada um dos seus públicos?', 'Quais modelos você pode preparar com antecedência?', 'Qual treinamento sua equipe precisa? e.g., diante das câmeras, falar em público, gerenciamento de crises nas redes sociais etc.' e 'Quem monitorará os novos desdobramentos da crise?' Inclua essas respostas em seu plano de comunicação, para que seja fácil encontrá-las e consultá-las quando forem mais necessárias.
13	Estabelecer feedback	Estabeleça um processo fácil para todos a quem você atribuiu uma tarefa possam documentar seus comentários.
14	Definir um cronograma para a atualização do plano de comunicação	Identifique a frequência com que você revisitará seu plano de comunicação. As alterações virão do feedback que você coletou na etapa anterior, bem como das mudanças do setor e de quaisquer novas datas importantes. Depois de decidir sua linha do tempo, certifique-se de adicionar uma nota sobre ela no próprio plano de comunicação.
15	Determinar métricas e marcos	Verifique novamente seus objetivos de negócios e metas de comunicação. Em seguida, identifique os dados que você precisa coletar para comparar com essas metas e objetivos. Defina a trajetória de crescimento que deseja ver com base em seus objetivos. Divida-o em marcos mensais e trimestrais.

Fonte: DiNardi (2019)

Os propósitos, as mensagens e os canais podem variar, mas a necessidade de manter interação com os stakeholders permanece. Depois de definir seu plano, combine-o com uma solução completa e robusta que ajude sua equipe a identificar rapidamente os stakeholders, direcioná-los para conteúdos relevantes, acompanhar resultados e implementar mudanças com eficiência.

Stakeholders devidamente identificados, categorizados, priorizados, comprometidos e informados, faremos uma pequena digressão em nossa análise. No capítulo seguinte, nos debruçaremos sobre quando, onde e como coletar os dados e informações sobre a IA. Esse ponto é essencial para demonstramos a factibilidade da implantação do processo proposto nesta obra e para ilustramos como viabilizar a segunda etapa (Registro das informações) do sistema de contabilidade. Retomaremos nossa saga no capítulo 7, "Quais tipos de informações serão coletados e quais divulgações serão feitas".

6 QUANDO, ONDE E COMO COLETAR?

Antes de prosseguirmos para nossa terceira questão normativa, “Quais tipos de informações serão coletadas e quais divulgações serão feitas?”, apresentaremos um pequeno estudo prático sobre quando, onde e como coletar essas informações sobre a IA. Começaremos por apresentar os conceitos de ex-ante e ex-post, para nos referirmos a um determinado fato envolvendo a IA. Em seguida, analisaremos a principal metodologia de *data mining* (i.e., mineração de dados) empregada na atualidade, CRoss-Industry Standard Process Model (CRISP-DM). Por fim, deslindaremos as principais arquiteturas de telemetria de software que podemos empregar para coletar dados ao longo do ciclo de vida da IA.

6.1 QUANDO COLETAR

Como visto no capítulo 2, *accountability* envolve duas responsabilidades ou deveres principais:

- Realizar certas ações, ou abster-se de, de acordo com as expectativas dos stakeholders e,
- Fornecer avaliação dessas ações para os interessados.

Também, optamos pela análise factual de qualidade (i.e., informação com relevância, confiabilidade, comparabilidade, verificabilidade e compreensibilidade) como método argumentativo para confeccionarmos nossos cômputos.

Quando a IA se envolve em algum incidente, dizemos que isso gera um fato pelo qual devemos realizar certas ações e prestar contas aos stakeholders. Todos os dados, análises, previsões, contenções e planejamentos que detemos, empregamos e planejamos sobre a IA até aquele momento são tidos como ex-ante, i.e., “antes do fato”. Outrossim, toda e qualquer análise após o fato será ex-post.

Talvez a forma mais natural de esclarecimento seja o exemplo. Em nosso contexto, nos concentramos em extrair espécimes representativos que compreen-

dem as relações internas, peculiaridades, causalidades e correlações externas de um modelo ou previsão para aclarar o porquê da ocorrência do fato, ou seja, porque o modelo se comporta dessa maneira em um determinado contexto. As fontes de nossas explicações serão os dados e o modelo.

Ao analisar os dados de treinamento (ex-ante), podemos entender as características e atributos (i.e., as variáveis correlacionadas) desses dados antes de iniciar o treinamento, identificar variáveis irrelevantes, descobrir ou verificar relacionamentos importantes que os modelos de aprendizado de máquina devem incorporar, e remover o viés subjacente aos dados antes de qualquer modelagem. Para uma análise ex-post, geralmente via dados telemétricos coletados, extraímos um conjunto cronológico de entradas e saídas, ou seja, a cadeia de eventos que levam a um fato, fornecendo uma interpretação bruta do que aconteceu e um conjunto factual para indagações posteriores sobre o modelo.

Analisando o modelo (ex-ante), podemos investigar questões sobre qualidade, segurança, proteção, robustez, transparência e assim por diante. O resultado dessas análises será crucial para sustentar um caso de previsibilidade, por exemplo. A utilização de técnicas de XAI como conjunto de ferramentas forenses (uso ex-post) permite explicar as deliberações e comportamentos da IA, em geral ou em situações específicas, durante a investigação. Em relação ao modelo, estaremos sempre preocupados em entender como ele funciona sob dois pontos de vista, global (holístico) e local (pontualmente), por meio de explicações, características e amostras.

6.2 ONDE COLETAR

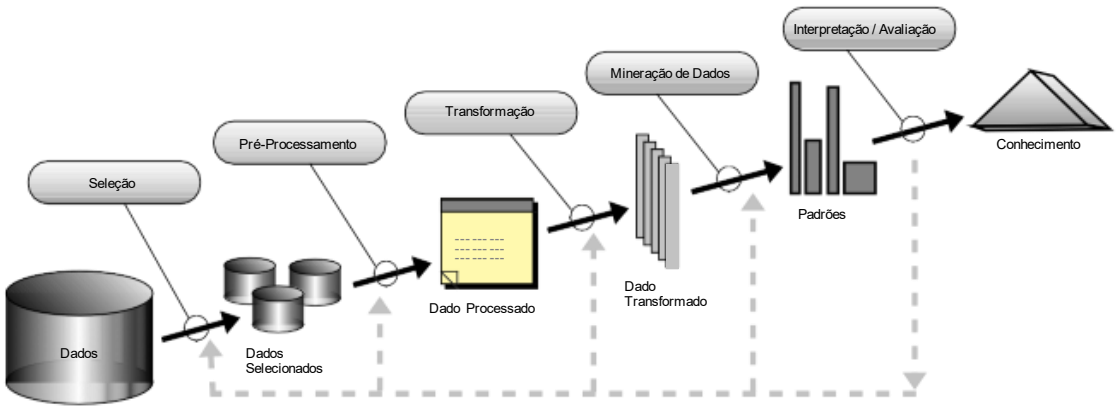
Dirimida a diferenciação de quando se dará a nossa investigação, é hora de perscrutar o ciclo de vida da IA de forma a viabilizar nossas inquirições ex-ante. Para tal, é importante examinar as duas principais metodologia de data mining.

A primeira metodologia, quer por sua importância histórica, quer pela facilidade em modificá-la para fins mais específicos, é o Knowledge Discovery in Databases (KDD). O KDD é um processo não trivial em várias etapas para extrair informações de uma base de dados.

Consiste no desenvolvimento de métodos e técnicas para dar sentido aos dados. O problema básico abordado pelo processo KDD é o de mapear dados de baixo nível em outras formas que podem ser mais compactas, abstratas ou úteis. No cerne do processo está a aplicação de métodos específicos de mineração de dados para descoberta e extração de padrões (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996). O pipeline KDD-Model consiste nas seguintes etapas:

1) compreensão do domínio do aplicativo,
2) seleção de dados,
3) limpeza e preparação de dados,
4) redução e transformação de dados,
5) seleção de métodos de mineração de dados adequados,
6) etapa de mineração de dados,
7) interpretação dos resultados encontrados,
8) desdobramento do conhecimento encontrado.

Figura 3 – Knowledge Discovery in Databases (KDD)



Fonte: Fayyad et al. (1996)

Acima a Figura 3 – Knowledge Discovery in Databases (KDD) fornece uma visão geral do pipeline de um projeto de mineração de dados. Ela contém as etapas (→) e suas principais saídas. A primeira e a última etapas foram suprimidas e as etapas 5 e 6 foram condensadas na etapa “Mineração de Dados”. A seguir, nos aprofundaremos no CRISP-DM.

6.2.1 CRoss-Industry Standard Process Model (CRISP-DM)

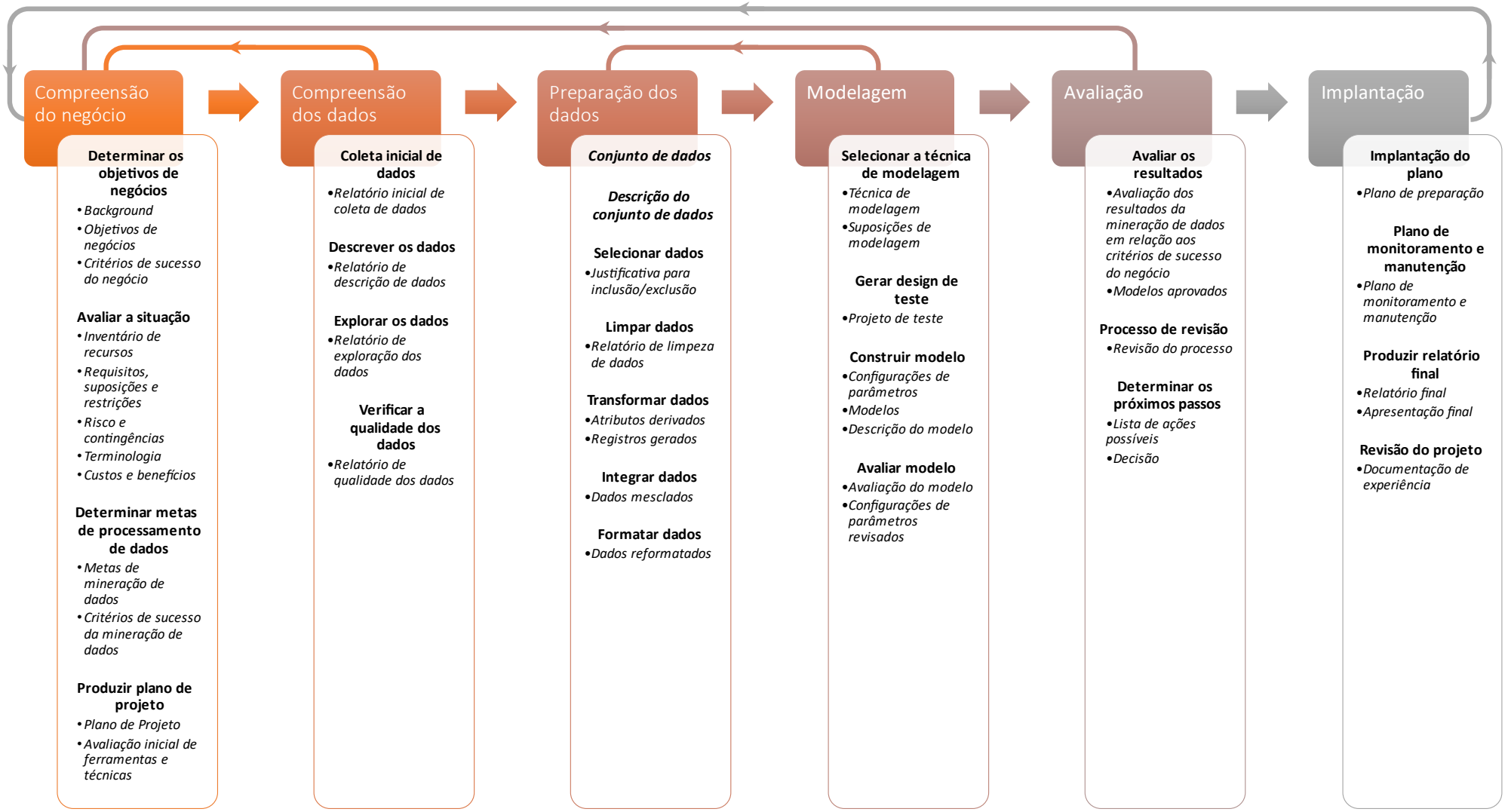
Embora o CRISP-DM seja uma metodologia de mineração de dados, a maioria dos desenvolvedores a segue com algumas adaptações (PIATETSKY-SHAPIRO, 2020). De acordo com o CRISP-DM, o processo envolve seis fases: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação (HARPER e PICKETT, 2006).

Na fase de compreensão do negócio, nos atemos a entender os objetivos e requisitos do projeto sob uma perspectiva do negócio. Na compreensão dos dados, identificamos problemas de qualidade de dados e detectamos subconjuntos interessantes para formar hipóteses sobre informações ocultas. A fase de preparação dos dados abrange todas as atividades para construir o conjunto de dados tratados (dados que serão alimentados no modelo) a partir dos dados brutos. Os dados tratados, geralmente, são desmembrados em três grupos com finalidades específicas: os dados de treinamento, validação e teste. As tarefas nessa fase incluem seleção de tabelas, registros e atributos, bem como transformação e limpeza de dados.

A fase seguinte, modelagem, também é conhecida como treinamento. É nessa fase que os dados de treinamento são aplicados às várias técnicas de modelagem. Os modelos resultantes são selecionados e seus parâmetros são calibrados para valores ótimos. Na fase de avaliação, se revisam os passos executados para a construção e conformação do modelo ante os objetivos de negócio desejados. A implantação, fase final, geralmente envolve a liberação do modelo para o cliente (CHAPMAN, CLINTON, *et al.*, 2000).

Abaixo a Figura 4 - Cross Industry Standard Process for Data Mining (CRISP-DM) fornece uma visão geral do ciclo de vida de um projeto de mineração de dados. Ela contém as fases de um projeto (retângulos), suas respectivas tarefas (**negrito**) e saídas (*itálico*). A sequência das fases não é rígida. Mover-se para frente e para trás entre as diferentes fases é sempre necessário. As setas indicam as dependências mais importantes e frequentes entre as fases. Essencialmente, podem existir relacionamentos entre quaisquer tarefas de mineração de dados, dependendo dos objetivos, histórico e interesse do usuário e, mais importante, dos dados. A seguir, abordaremos cada uma das fases, suas tarefas e respectivas saídas (↻).

Figura 4 - Cross Industry Standard Process for Data Mining (CRISP-DM)
Tarefas genéricas (**negrito**) e saídas (itálico), as setas indicam as dependências mais importantes e frequentes entre as fases.



Fonte: Chapman et al. (2000)

6.2.1.1 Compreensão do negócio

Esta fase inicial se concentra na compreensão dos objetivos e requisitos do projeto de uma perspectiva de negócios, convertendo esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para atingir os objetivos. Esta fase é composta por 4 tarefas: determinar os objetivos de negócio, avaliar a situação, determinar metas de mineração de dados e produzir um plano de projeto.

Determinar os objetivos de negócios – o primeiro objetivo do analista de dados é entender completamente, de uma perspectiva de negócios, o que o cliente realmente deseja realizar.

- *Background* – registra as informações que são conhecidas pela organização no início do projeto sobre o tema do projeto.
- *Objetivos de negócios* – descreve o objetivo principal do cliente, de uma perspectiva de negócios.
- *Critérios de sucesso do negócio* – descreve os critérios para um resultado bem-sucedido ou útil para o projeto do ponto de vista do negócio.

Avaliar a situação – esta tarefa envolve a descoberta de fatos mais detalhados sobre todos os recursos, restrições, suposições e outros fatores que devem ser considerados na determinação do objetivo da análise de dados e do plano do projeto.

- *Inventário de recursos* – lista os recursos disponíveis para o projeto, incluindo: pessoal (especialistas em negócios, especialistas em dados, suporte técnico, pessoal de mineração de dados), dados (extrações, acesso a dados operacionais ou armazenados), recursos computacionais (plataformas de hardware) e software (ferramentas de mineração de dados, outro software relevante).
- *Requisitos, suposições e restrições* - lista todos os requisitos do projeto, incluindo cronograma de conclusão, compreensibilidade e

qualidade dos resultados e segurança, bem como questões legais. Lista as suposições feitas e as restrições do projeto.

- *Riscos e contingências* – lista os riscos ou eventos que podem ocorrer para atrasar o projeto ou fazer com que ele falhe.
- *Terminologia* – compila um glossário de terminologia relevante para o projeto.
- *Custos e benefícios* – uma análise de custo-benefício para o projeto, que compara os custos do projeto com o benefício potencial para o negócio se for bem-sucedido.

Determinar metas de mineração de dados – uma meta de negócios declara os objetivos na terminologia de negócios. Uma meta de mineração de dados declara os objetivos do projeto em termos técnicos.

- *Metas de mineração de dados* – descreve as saídas pretendidas do projeto que permitem a realização dos objetivos de negócios.
- *Critérios de sucesso da mineração de dados* – define os critérios para um resultado bem-sucedido do projeto em termos técnicos.

Produzir plano de projeto – descreve o plano pretendido para atingir as metas de mineração de dados e alcançar os objetivos do negócio.

- *Plano de projeto* – lista as etapas a serem executadas no projeto, juntamente com duração, recursos necessários, entradas, saídas e dependências.
- *Avaliação inicial de ferramentas e técnicas* – ao final da primeira fase, o projeto também realiza uma avaliação inicial de ferramentas e técnicas.

6.2.1.2 Compreensão dos dados

A fase de compreensão de dados começa com uma coleta inicial de dados e prossegue com atividades para se familiarizar, identificar problemas de qualidade e descobrir os primeiros insights sobre os dados, além de detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas. Esta fase é composta por 4 tarefas: coletar, descrever, explorar e verificar a qualidade dos dados coletados.

Coleta inicial de dados – adquire (ou acessa) os dados listados nos recursos do projeto. Essa coleta inicial inclui o carregamento de dados, se necessário, para a etapa de ‘compreensão dos dados’.

- *Relatório inicial de coleta de dados* – lista o conjunto de dados (ou conjuntos de dados) adquiridos, juntamente com suas localizações dentro do projeto, os métodos usados para adquiri-los e quaisquer problemas encontrados.

Descrever os dados – examina as propriedades “brutas” ou “superficiais” dos dados adquiridos e relata os resultados.

- *Relatório de descrição dos dados* – Descreve os dados que foram adquiridos, incluindo: o formato dos dados, a quantidade de dados (e.g., número de registros e campos em cada tabela), as identidades dos campos e quaisquer outras características superficiais dos dados que foram descobertas.

Explorar os dados – esta tarefa aborda as questões de mineração de dados, que podem ser desenvolvidas usando consultas, visualizações e relatórios.

- *Relatório de exploração dos dados* – descreve os resultados desta tarefa, incluindo as primeiras descobertas ou hipóteses iniciais e seus impactos no restante do projeto.

Verificar a qualidade dos dados – examina a qualidade dos dados, abordando questões como: valores ausentes ou integridade, correção e erros de dados.

- *Relatório de qualidade dos dados* – lista os resultados da verificação da qualidade dos dados; se problemas de qualidade existirem, listar possíveis soluções.

6.2.1.3 Preparação dos dados

A fase de preparação dos dados abrange todas as atividades para construir o conjunto de dados tratados (dados que serão alimentados na(s) ferramenta(s) de modelagem) a partir dos dados brutos iniciais. As tarefas da fase de preparação dos dados provavelmente serão executadas várias vezes e não numa ordem prescrita. As tarefas incluem seleção de tabela, registro e atributo, bem como transformação e limpeza de dados para conformar as demandas específicas das ferramentas de modelagem. Esta fase é marcada pela confecção do conjunto de dados tratados e sua respectiva descrição e composta por 5 tarefas: selecionar, limpar, construir, integrar e formatar os dados.

- *Conjunto de dados tratados* – este é o conjunto de dados (geralmente, dados de treinamento, validação e teste) produzidos pela fase de preparação de dados, que serão usados para modelagem ou o trabalho de análise principal do projeto.
- *Descrição do conjunto de dados* – descreve o conjunto de dados tratados que será usado para modelagem ou o trabalho de análise principal do projeto.

Selecionar dados – decide sobre os dados a serem usados para análise. Os critérios incluem relevância para as metas de mineração de dados, qualidade e restrições técnicas, como limites de volume ou tipos de dados.

- *Justificativa para inclusão/exclusão* – lista os dados a serem incluídos/excluídos e as razões para essas decisões.

Limpar dados – aumenta a qualidade dos dados para o nível exigido pelas técnicas de análise selecionadas.

- *Relatório de limpeza de dados* – descreve quais decisões e ações foram tomadas para resolver os problemas de qualidade de dados relatados durante a tarefa de verificação de qualidade de dados da fase de compreensão de dados.

Transformar dados – esta tarefa inclui operações construtivas de preparação de dados, como a produção de atributos derivados, novos registros completos ou valores transformados para atributos existentes.

- *Atributos derivados* – descreve os novos atributos construídos a partir de um ou mais atributos existentes no mesmo registro.
- *Registros gerados* – descreve a criação de registros completamente novos.

Integrar dados - esta tarefa inclui o emprego de métodos pelos quais as informações são combinadas de várias tabelas ou registros para criar novos registros ou valores. Mesclar tabelas refere-se à junção de duas ou mais tabelas que possuem informações diferentes sobre os mesmos objetos. Dados mesclados também cobrem agregações. Agregação refere-se a operações em que novos valores são calculados resumindo informações de vários registros e/ou tabelas

- *Dados mesclados* – contêm os dados que foram mesclados ou agregados.

Formatar dados – as transformações de formatação referem-se principalmente a modificações sintáticas feitas nos dados que não alteram seu significado, mas podem ser exigidas pela ferramenta de modelagem.

- *Dados reformatados* – dados editados em determinado formato padrão. Abrange modificações como alterações sintáticas feitas para satisfazer os requisitos da ferramenta de modelagem específica, alterações na ordem dos atributos e randomização.

6.2.1.4 Modelagem

Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados para valores ótimos. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos específicos na formatação dos dados. Portanto, muitas vezes é necessário voltar à fase de preparação de dados. Esta fase é composta por 4 tarefas: selecionar a técnica de modelagem, gerar o design de teste, construir o modelo e avaliar o modelo.

Selecionar a técnica de modelagem – como primeira etapa na modelagem, seleciona a técnica de modelagem real que será usada, ou seja, a técnica de modelagem específica. Se várias técnicas forem aplicadas, execute esta tarefa para cada método separadamente.

- *Técnica de modelagem* – documenta a técnica de modelagem que será usada.
- *Suposições de modelagem* – muitas técnicas de modelagem fazem suposições específicas sobre os dados, portanto, registra todas as suposições feitas.

Gerar design de teste – antes de realmente construirmos um modelo, precisamos gerar um procedimento ou mecanismo para testar a qualidade e validade do modelo. Portanto, normalmente separamos o conjunto de dados em conjunto de treinamento e teste, construímos o modelo no conjunto de treinamento e estimamos sua qualidade no conjunto de teste separado.

- *Projeto de teste* – descreve o plano pretendido para treinar, testar e avaliar os modelos. Um componente primário do plano é decidir como dividir o conjunto de dados disponível em dados de treinamento, dados de teste e dados de validação.

Construir modelo – executa a ferramenta de modelagem no conjunto de dados preparado para criar um ou mais modelos.

- *Configurações de parâmetros* – com qualquer ferramenta de modelagem, muitas vezes há muitos parâmetros que podem ser ajustados. Lista os parâmetros e seus valores escolhidos, juntamente com a justificativa para a escolha das configurações dos parâmetros.
- *Modelos* – esses são os modelos reais (i.e., a IA propriamente dita) produzidos pela ferramenta de modelagem, não um relatório.
- *Descrição do modelo* – descreve o modelo resultante. Relata as interpretações do modelo (i.e., parâmetros, métricas, limitações etc.) e documenta quaisquer dificuldades encontradas (e.g., vieses, discrepâncias, erros etc.).

Avaliar modelo – o engenheiro de mineração de dados interpreta os modelos de acordo com seu conhecimento de domínio, os critérios de sucesso da mineração de dados e do projeto de teste. Na medida do possível, ele também leva em consideração os objetivos de negócio e os critérios de sucesso do negócio. Na maioria dos projetos de mineração de dados, o engenheiro aplica uma única técnica mais de uma vez ou gera resultados com diferentes técnicas alternativas. Nesta tarefa, ele também compara todos os resultados de acordo com os critérios de avaliação.

- *Avaliação do modelo* – resume os resultados desta tarefa, lista as qualidades dos modelos gerados e ranqueia sua qualidade em relação aos outros.
- *Configurações de parâmetros revisadas* – de acordo com a avaliação do modelo, o engenheiro revisa as configurações de parâmetro e ajustes, até que encontra o(s) melhor(es) modelo(s), para a próxima execução da tarefa **construir modelo**. Documenta todas essas revisões e avaliações.

6.2.1.5 Avaliação

Nesta fase do projeto, você construiu um modelo (ou modelos) que parece ter alta qualidade do ponto de vista da análise de dados. Antes de prosseguir para a implantação final do modelo, é importante avaliar mais detalhadamente o modelo e revisar as etapas executadas para construir o modelo de forma a garantir que ele atinja adequadamente os objetivos de negócios.

Um dos principais objetivos é determinar se há alguma questão comercial importante que não foi suficientemente considerada. No final desta fase, deve-se decidir sobre a utilização dos resultados da mineração de dados. Esta fase é composta por 3 tarefas: avaliar os resultados, processo de revisão e determinar os próximos passos.

Avaliar os resultados – esta etapa avalia o grau de aderência do modelo aos objetivos de negócios e procura determinar se há algum motivo comercial para que esse modelo seja invalidado. Outra opção de avaliação é testar o(s) modelo(s) em versões BETA do aplicativo final, se as restrições de tempo e orçamento permitirem. Além disso, a avaliação também estima outros resultados gerados, como todas as outras descobertas que não estão necessariamente relacionadas aos objetivos de negócios originais, mas podem revelar desafios adicionais, informações ou dicas para direções futuras.

- *Avaliação dos resultados da mineração de dados em relação aos critérios de sucesso do negócio* – resume os resultados da avaliação em termos de critérios de sucesso do negócio, incluindo uma declaração final se o projeto cumpre os objetivos iniciais do negócio.
- *Modelos aprovados* – após a avaliação do modelo com relação aos critérios de sucesso do negócio, os modelos gerados que atendem aos critérios selecionados tornam-se modelos aprovados

Processo de revisão – agora é apropriado fazer uma revisão mais completa do processo para determinar se há algum fator ou tarefa importante

que tenha sido negligenciado de alguma forma. Esta revisão também abrange questões de garantia de qualidade.

- *Revisão do processo* – resume a revisão do processo e destaca as atividades que foram perdidas e/ou devem ser repetidas.

Determinar os próximos passos – o gerente precisa decidir se deve concluir este projeto e passar para a implantação, se apropriado, ou se deve iniciar outras iterações ou configurar novos projetos de mineração de dados. Esta tarefa inclui análises de recursos remanescentes e orçamento que influenciam as decisões.

- *Lista de ações possíveis* – lista as possíveis ações futuras junto com as razões a favor e contra cada opção.
- *Decisão* – descreve a decisão sobre como proceder com o projeto.

6.2.1.6 Implantação

A criação do modelo geralmente não é o fim do projeto. Mesmo que o objetivo do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido precisará ser organizado e apresentado de forma que o cliente possa utilizá-lo. Muitas vezes, envolve a aplicação de modelos dinâmicos nos processos de tomada de decisão de uma organização, por exemplo, na personalização em tempo real de páginas da Web ou pontuação repetida de bancos de dados de marketing. No entanto, dependendo dos requisitos, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados replicável em toda a empresa. Em muitos casos, é o cliente, não o analista de dados, quem realiza as etapas de implantação. Porém, mesmo que o analista não vá realizar o trabalho de implantação, é importante que o cliente entenda de antemão quais ações precisam ser realizadas para de fato fazer uso dos modelos criados. Esta fase

é composta por 4 tarefas: implantação do plano, plano de monitoramento e manutenção, produzir o relatório final e revisar o projeto.

Implantação do plano – para implantar o(s) resultado(s) da mineração de dados no negócio, esta tarefa pega os resultados da avaliação e conclui uma estratégia de implantação. Se um procedimento geral foi identificado para criar o(s) modelo(s) relevante(s), esse procedimento é documentado aqui para implantação posterior.

- *Plano de preparação* – resume a estratégia de implantação, incluindo as etapas necessárias e como executá-las.

Plano de monitoramento e manutenção – o monitoramento e a manutenção são questões importantes se o resultado da mineração de dados se tornar parte do dia a dia dos negócios e de seu ambiente, ajudando a evitar períodos desnecessariamente longos de uso incorreto dos resultados da mineração de dados. Para monitorar a implantação do(s) resultado(s), o gestor precisa de um planejamento detalhado do processo de monitoramento, ou seja, levando em consideração o tipo específico de implantação.

- *Monitoramento e plano de manutenção* – resuma a estratégia de monitoramento e manutenção, incluindo as etapas necessárias e como executá-las.

Produzir relatório final – dependendo do plano de implantação, este relatório pode ser apenas um resumo do projeto e suas experiências ou pode ser uma apresentação final e abrangente do(s) resultado(s) da mineração de dados.

- *Relatório final* – este é o relatório final que inclui todos os resultados anteriores e resume e organiza os resultados.
- *Apresentação final* – também haverá frequentemente uma reunião na conclusão do projeto em que os resultados são apresentados verbalmente ao cliente.

Revisão do projeto – Avalia o que deu certo e o que deu errado, o que foi bem-feito e o que precisa ser melhorado.

- *Documentação da experiência* – resume conhecimentos importantes feitos durante o projeto. Por exemplo, armadilhas, abordagens enganosas, dicas para selecionar as técnicas de mineração de dados mais adequadas em situações semelhantes, quaisquer relatórios que tenham sido escritos por membros individuais do projeto durante as fases do projeto e suas tarefas podem fazer parte desta documentação.

Ao compararmos as fases do CRISP-DM com as do KDD fica claro perceber a semelhança entre os dois. O modelo CRISP-DM combina as etapas 3 e 4 do KDD na fase de preparação dos dados e as etapas 5 e 6 na fase de modelagem.

Desde o lançamento do CRISP-DM, no ano 2000, até hoje, inúmeros modelos foram surgindo para acomodar novas tecnologias e conceitos, e.g., *edge computing*, *Meta-data*, *data quality*, *data security* etc. Contudo, nenhum desses novos modelos difere em essência do CRISP-DM.

A supressão de certas saídas, a adaptação de certas tarefas e a combinação de certas fases se farão presentes no processo de conformação à novas tecnologias e realidades. Entretanto, mesmo que lancemos mão da mais hodierna metodologia ágil de data mining, ela deve suportar, além das supracitadas propriedades, sua capacidade de produzir uma análise factual de qualidade, i.e., suportar seu processo de *accountability*.

Uma metodologia como o CRISP-DM provê uma série de artefatos que nos permite acompanhar, mensurar e desenvolver de maneira satisfatória o pacto. Essas saídas estão em conformidade com as nossas duas primeiras questões normativas. Ou seja, elas nos proporcionam dados suficientes para atendermos às demandas dos diversos stakeholders e, quando não presentes no ciclo original, podemos facilmente adaptá-lo para obtê-las (como veremos no capítulo seguinte).

O Plano de monitoramento e manutenção do CRISP-DM é deveras importante quando o resultado da IA se tornar parte do dia a dia dos negócios, e afeta diversos stakeholders. Uma preparação cuidadosa de uma estratégia de

manutenção e prevenção, ajuda a evitar uma série de riscos, que variam de períodos desnecessariamente longos de uso incorreto dos resultados até acidentes e fatalidades. Para monitorar a implantação do(s) resultado(s) da mineração de dados, o projeto precisa de um plano detalhado sobre o processo de monitoramento.

Um bom Plano de monitoramento e manutenção leva em consideração o tipo específico de implantação, e.g., um aplicativo de reconhecimento facial para destravar o celular (software) ou um sistema de navegabilidade para carro autônomo (uma série de subsistemas produzidos por terceiros em hardware, software e serviços). Igualmente, nos permite uma análise factual de qualidade dos eventos pós-implantação da IA (ex-post).

Algumas das atividades a serem contempladas pelo plano incluem: verificar os aspectos dinâmicos (i.e., o que pode mudar no ambiente?), a precisão (i.e., como a precisão será monitorada?) e a vida útil (i.e., quando o resultado ou modelo não deve mais ser usado?). Precisamos, também, identificar critérios (validade, limite de precisão, novos dados, mudança no domínio do aplicativo etc.) e o que deve acontecer se o modelo ou resultado não puder mais ser usado? (i.e., atualizar o modelo, configurar novo projeto de mineração de dados etc.) e se os objetivos de negócios do uso do modelo mudarão com o tempo (i.e., deve-se documentar totalmente o problema inicial que o modelo estava tentando resolver para se verificar sua atual aderência).

A seguir, estudaremos como automatizar uma boa parte da coleta de dados sobre a IA durante seu ciclo de vida. Através da telemetria, podemos colher, limpar, preparar, reduzir, transformar, minerar e interpretar os dados da IA tanto ex-ante quanto ex-post.

6.3 COMO COLETAR

Houve um tempo em que o monitoramento simples de telemetria era tudo o que as empresas precisavam para processar erros e resolver problemas do sistema. Telemetria são os dados que os sistemas (em produção, operação, suporte etc.) emitem para fornecer feedback sobre o que está acontecendo (i.e., fornecem um contexto sobre a produção, operação ou falha do sistema).

Porém, os tempos mudaram. Inteligência de negócios, camadas de acesso a dados e serviços de aplicativos agora são frequentemente implantados de forma independente, tornando cada vez mais difícil monitorar onde e quando as interrupções acontecem.

6.3.1 Telemetria

Telemetria é a coleta in situ de medições ou outros dados em pontos remotos e sua transmissão automática, para equipamentos receptores, para monitoramento (NASA, 1987). O início da telemetria industrial data da era do vapor, mas a telemetria em si, apesar de origem incerta, remonta as tábuas astrológicas e de maré (KOPP, 2002).

O advento da Segunda Guerra Mundial impulsionou o desenvolvimento industrial e a utilização da aviação como meio de transporte de massa. Daí em diante muitos desses equipamentos telemétricos tornaram-se comercialmente viáveis, obrigatórios e conhecidos do público (KOPP, 2002). Contudo, um sistema telemétrico hodierno não provê apenas dados, ele guarnece observabilidade.

Observabilidade não é apenas sobre ferramentas e painéis. Trata-se de construir uma base sólida de bons fundamentos de registros e métricas. Em uma arquitetura de observabilidade é possível identificar a origem do problema apesar de sua possível obscuridade, i.e., você pode inferir o estado interno de

um sistema com base em suas saídas. Como não podemos prever tudo o que queremos saber sobre nosso sistema ex-ante, queremos rastrear dados suficientes para garantir que possamos analisar problemas de diferentes ângulos e diferentes agregados quando os problemas inevitavelmente ocorrerem ex-post.

Existem três pilares principais que a observabilidade abrange: métricas, logs e rastreamentos. Essas três soluções combinadas, devem fornecer uma visão holística para ajudar a diagnosticar o que está acontecendo dentro do sistema e eventuais causadores.

> **Métricas** – geralmente agregam dados numéricos medidos durante um certo intervalo de tempo. As métricas consistem em um conjunto de atributos como nome, carimbo de data/hora, valor e rótulos.

Normalmente, usamos métricas para determinar a integridade do sistema, i.e., descrever o status dos recursos. Depois que as métricas do sistema são definidas, você também pode adicionar métricas personalizadas, que fornecem informações relevantes sobre o negócio ou domínio, ao equipar seu código com bibliotecas que emitem parâmetros customizadas para obter melhores insights sobre o sistema.

Existem duas classes de métricas: métricas simples (e.g., contadores) e eventos complexos (geralmente métricas de negócios ou domínio). Alguns dos tipos mais comuns de métricas, incluem:

- **Medidores** – representam aferições realizadas num determinado ponto no tempo. Por exemplo, temperatura, velocidade de conexão etc.
- **Contadores** – medem a quantidade de eventos que ocorreram. Por exemplo, você pode contar o número de erros de classificação ou o número de acesso por segundo.
- **Histogramas** –são utilizados principalmente para demonstrar dados colhidos em estatísticas, e.g., medidores ou contadores. É basicamente um gráfico usado para análise de dados, formado por colunas que indicam a frequência dos dados obtidos, o que facilita a observação dos valores.

- **Constantes** – são usadas para informações que não mudam durante o processo. Isso pode indicar um número de versão do aplicativo ou um vetor de inicialização.

Com as métricas, você tem o potencial de medir qualquer coisa que ocorra em seu sistema. No entanto, há uma ressalva. Ao monitorar métricas de controle, i.e., que acionam um alerta, estamos analisando problemas conhecidos que ocorreram no passado.

Idealmente, sabemos quais são as implicações de uma métrica ultrapassar o limite e sabemos como corrigi-la. Outras vezes, o que acontece é que nossas métricas estão nos dizendo que o sistema está saudável, mas alguns usuários continuam reclamando que o sistema está inoperante. Portanto, as métricas por si só não são suficientes. Você precisa ter mais contexto.

Logs – fornecem dados textuais sobre eventos que ocorrem em seu sistema. Alguns logs podem fornecer informações cruciais para resolver problemas. Eles são essenciais para entender as atividades de sistemas complexos, principalmente no caso de aplicações com pouca interação do usuário (e.g., a IA).

Vale a pena agregar logs porque eles fornecem contexto para recriar e investigar problemas. Esse registro pode ser utilizado para restabelecer o estado original de um sistema ou para que se conheça o seu comportamento no passado. Ademais, possuem grande importância para o Direito da Tecnologia da Informação como meio de prova digital.

Rastreamento (ou logs de rastreamento) – capturam informações sobre o ambiente operacional (i.e., todas as instruções executadas e os bytes de memória acessados) quando o software falha ao operar como pretendido.

Tanto os logs quanto as métricas podem estar relacionados a eventos específicos que ocorrem no sistema. Contudo, essas telemetrias não fornecem a capacidade de rastrear uma transação, ou operação específica, até adicionarmos o rastreamento. Por exemplo, se você deseja acompanhar uma deter-

minada transação que falhou, você pode examinar os rastreamentos e vincular métricas, erros e logs relevantes para mostrar a execução da referida transação (i.e., uma trilha auditável através do código e da arquitetura e seu contexto de execução).

Aqui é importante diferenciar os conceitos de observabilidade e monitoramento. Embora os conceitos pareçam semelhantes, eles diferem quanto a finalidade.

O monitoramento é uma ferramenta, ou uma solução técnica, que permite as equipes observarem e entenderem o estado de seus sistemas. O monitoramento é baseado na coleta de conjuntos predefinidos de métricas ou logs.

A observabilidade é uma ferramenta, ou uma solução técnica, que permite as equipes depurarem ativamente o sistema. A observabilidade é baseada na exploração de propriedades e padrões não definidos previamente. Idealmente, os dois conceitos devem ser suportados na confecção de um sistema de telemetria.

6.3.1.1 Sistemas de telemetria

São os sistemas que manipulam, transformam, armazenam e apresentam dados telemétricos. Sua principal função é agregar e transformar os dados telemétricos das diversas fontes e apresentá-los para a tomada de decisões. Os quatro principais estilos de telemetria em uso hoje são:

Registro Centralizado (logging) – Esse foi o primeiro estilo de telemetria criado e data do início dos anos 1980. Consiste em agrupar os registros gerados (dados textuais) pelos sistemas de produção para um local central onde podem ser consultados. Suporta não apenas telemetria de software, mas também telemetria de hardware.

Métricas – Esse estilo surgiu no início dos anos 2010. Se concentra em números (contadores, cronômetros, taxas etc.), em vez de texto, para

descrever o que está acontecendo. As métricas permitem que prazos muito maiores sejam mantidos on-line e pesquisáveis quando comparados com o registro centralizado.

Rastreamento Distribuído – Esse estilo surgiu no final da década de 2010. O rastreamento distribuído é uma união dos estilos métrica e *logging* (i.e., a cardinalidade[51] do *logging* com o poder analítico das métricas) e concentra-se diretamente no rastreamento de eventos em múltiplos componentes de um sistema distribuído. O rastreamento permite adicionar fluxos de execução inteiros ao contexto apresentado quando se investiga o que aconteceu durante uma execução. O *logging* exemplifica eventos específicos conforme eles acontecem, com alguns atributos para fornecer contexto sobre o evento. Já as métricas fornecem uma visão geral ampla do desempenho do sistema. O rastreamento usa o contexto extra do *logging* para criar uma exibição encadeada das execuções registradas. Essa visualização torna mais fácil para as pessoas isolarem rapidamente pontos interessantes para investigar.

Gerenciamento de Eventos de Segurança da Informação (GESI) – Esse estilo surgiu em meados dos anos 1990. É um sistema de telemetria especializado desenvolvido para atender as demandas específicas das equipes de segurança e conformidade (i.e., rastrear login e logout, uso de privilégios, acesso a dados sensíveis etc.). Como esses requisitos são tão comuns, e rastreá-los e correlacioná-los posteriormente é complexo, eles deram origem a um estilo de telemetria separado. Devido à complexidade da tarefa, os GESI quase sempre são softwares pagos e você só terá que conectar fontes de telemetria ao sistema que sabe como interpretar esses dados.

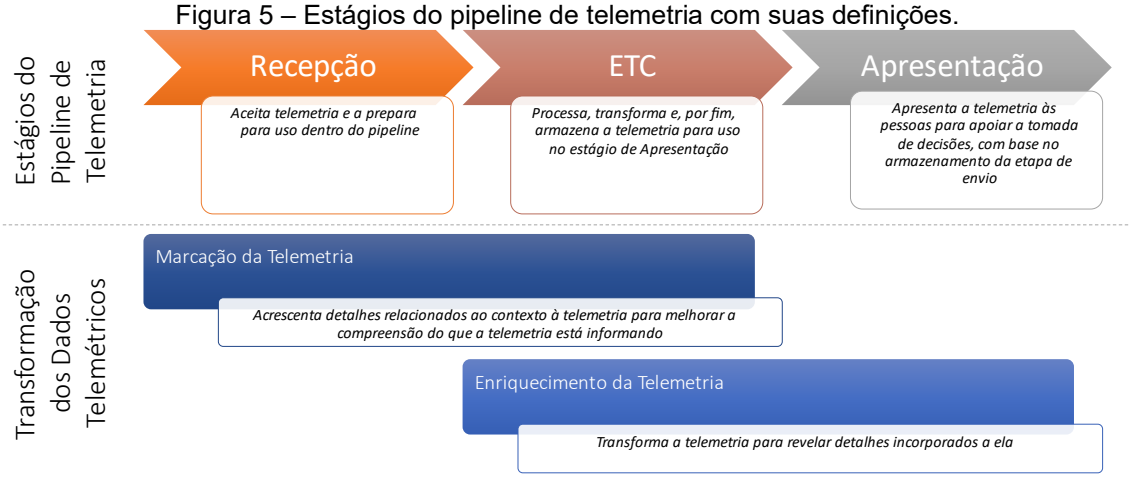
---

[51] Cardinalidade é o número de combinações únicas que os campos podem produzir o que afeta significativamente o desempenho da pesquisa. Por exemplo, se você tiver campos A e B, onde A tem dois valores possíveis e B tem três valores possíveis, a cardinalidade desse índice é A * B = 6.

6.3.1.1.1 Arquitetura

Cada um dos sistemas de telemetria descritos acima segue a mesma arquitetura geral, ilustrado na Figura 5.

Figura 5 – Estágios do pipeline de telemetria com suas definições.



Fonte: Riedesel (2021)

No primeiro estágio, Recepção, todos os dados telemétricos auferidos são agregados e preparados para uso dentro do seu pipeline de ETL[52] de dados telemétricos, estágio seguinte ETC. Os sistemas que recebem as emissões do estágio anterior processam e transformam a telemetria para prepará-la para o armazenamento. Finalmente, no estágio Apresentação, os dados telemétricos armazenados são analisados e visualizados, subsidiando decisões informadas (e.g., relatórios ou dashboards) para os stakeholders interessados. Ao longo do pipeline, a telemetria é marcada com detalhes relacionados ao contexto e, em seguida, enriquecida para extrair detalhes codificados (RIEDESEL, 2021).

A essa altura, o leitor mais familiarizado com o KDD deve ter notada a semelhanças entre algumas de suas fases e os estágios do pipeline de telemetria. Notadamente, o estágio de Recepção se assemelha as etapas de Seleção

[52] ETL – Extract, Transform, Load, ou ETC – Extrair, Transformar, Carregar (em português), é o processo de carga de dados, utilizando em integração de sistemas. O processo consiste na extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e, por fim, a carga dos dados em um Data Mart ou um Data Warehouse.

e e/ou Pré-processamento do KDD (a depender da arquitetura que empreguemos); ETC as etapas de Pré-processamento e Transformação e, por fim, o estágio de Apresentação equivale as etapas de Mineração de Dados e Interpretação / Avaliação.

Ademais, cada um dos três estágios pode ser implementado num único aparelho ou por uma miríade de sistemas que concorrem para o perfazimento das tarefas da etapa. Dessa forma, a arquitetura, apesar de simples, é robusta o suficiente para acomodar os diversos estilos e peculiaridades de implementação.

6.3.1.1.1.1 Recepção

A principal função desse estágio é aceitar a telemetria enviada pelas diversas fontes e prepará-la para uso dentro do pipeline. Aqui, a grande diferença entre métricas e logs é o formato dos dados: números e alguns detalhes extras no caso de métricas versus strings para logs.

Geralmente, o estágio de Recepção é implementado como um receptor (e.g., função, servidor etc.) que recebe dos diversos emissores (código, hardware, *loggers* etc.) a telemetria referente ao módulo assistido, e em seguida registra, agrega, formata e grava/transmite os dados para o estágio ETC.

O estágio de Recepção é o melhor lugar para você aplicar marcações (i.e., descrições sobre o contexto do evento emitido) porque o mais próximo ao contexto do que está sendo emitido. O melhor lugar para aplicar a marcação é dentro do próprio sistema de produção. Marcadores de contexto úteis podem incluir a classe, o método e a chamada de função em que o código estava quando a telemetria foi emitida, bem como parâmetros úteis ou interessantes (desde que não sejam de privados ou relacionados à saúde).

6.3.1.1.1.2 Extrair Transformar e Carregar (ETC)

Como vimos anteriormente, a etapa de Recepção envia os dados telemétricos para serem armazenados numa base de dados. A primeira tarefa da etapa de ETC é, portanto, acessar esses dados. Além da função extrair, a etapa de ETC desempenha duas outras funções principais: transformar e carregar.

Extrair – Durante a extração, o ETC identifica os dados e os reproduz de suas origens (mover a telemetria pode ser bastante complexo, i.e., envolver sistemas pertencentes e controlados por diferentes equipes.), de forma que possa transportá-los para o armazenamento de dados de destino.

Transformar – Como os dados extraídos são brutos em sua forma original, eles precisam ser preparados (i.e., mapeados e transformados) para eventual armazenamento. No processo de transformação, o ETC valida, autentica, remove duplicatas e/ou agrega as telemetrias de forma que tornam os dados resultantes confiáveis e consultáveis. Nessa etapa também ocorre a marcação (i.e., adicionar contexto) e enriquecimento (i.e., destacar detalhes) da telemetria.

Carregar – O ETC move os dados transformados para o armazenamento de dados de destino. Esta etapa pode implicar o carregamento inicial de todos os dados de origem ou pode ser o carregamento de alterações incrementais nos dados de origem. Você pode carregar os dados em tempo real ou em lotes programados.

O estágio ETC precisa não apenas extrair, transformar e carregar a telemetria, mas também, dar suporte aos formatos que seus componentes utilizam durante a marcação e o enriquecimento que ocorre durante o estágio. Um equívoco comum é crer que apenas nesse estágio ocorra toda a marcação e enriquecimento. Embora o estágio ETC execute boa parte de ambos, a marcação pode ocorrer durante os estágios Recepção e ETC, e o enriquecimento pode ocorrer nos estágios ETC e Apresentação.

O estágio ETC é especialmente importante para a marcação quando o estágio Recepção não está sob seu controle, e.g., com emissões de hardware ou emissões provenientes de outras plataformas. Podemos aplicar métodos de decodificação e extração de significado à telemetria para criar campos que melhoram a capacidade de pesquisa. Nesses casos, a adição de marcação, relacionada ao contexto que gerou à telemetria, aumenta muito a capacidade de alguém encontrar eventos relacionados, conforme Figura 5 – Estágios do pipeline de telemetria com suas definições..

A outra função principal do estágio ETC é reformatar a telemetria para armazenamento e posterior uso pelo estágio de Apresentação. Essas operações de conversão de tipos melhoram a capacidade do sistema de Apresentação de analisar os dados.

6.3.1.1.1.3 Apresentação

O estágio de apresentação é a última etapa do pipeline de telemetria e aquele que a maioria das pessoas na organização (e fora dela) usa para interagir com a telemetria. De muitas maneiras, essa etapa é a “cara” do ecossistema de telemetria. Como o estágio de apresentação é a principal forma de consumo de telemetria, é importante ter os sistemas de apresentação corretos para as decisões que precisam ser tomadas.

O trabalho dos sistemas desse estágio é filtrar, transformar, agregar e, opcionalmente, fornecer análises complexas sobre os dados, tudo para produzir tabelas, gráficos, painéis e relatórios de que as pessoas prescindem para tomar decisões informadas. Quando uma tabela, gráfico, painel ou relatório é solicitado, a maior parte do trabalho de transformação ocorre em tempo real. Por esse motivo, a engenharia por trás desses sistemas está entre as mais complexas em todo o pipeline de telemetria, especialmente no caso de rastreamento distribuído.

Quando todos os três Pilares de Observabilidade (logs, métricas e rastreamentos) são empregados em sistemas de apresentação adequados para suas tarefas, sua organização estará em melhor posição para aprender como seu sistema está operando (i.e., observabilidade). Cada vez mais, os sistemas de apresentação estão ganhando a capacidade de definir alarmes em tempo real, que são usados para notificar eventos críticos à medida que eles acontecem (i.e., monitoramento).

O enriquecimento no estágio de apresentação é menos sobre a modificação da telemetria para melhorar o desempenho e mais sobre a criação de visualizações úteis para os stakeholders. Painéis e gráficos são exibições visuais tanto quanto tabelas com colunas selecionadas. Linhas de tendência em gráficos e colunas de previsão para relatórios criados a partir da telemetria são outra forma de enriquecimento que os sistemas de apresentação produzem.

Sendo o último estágio no pipeline de telemetria, o estágio de Apresentação usa a telemetria que foi marcada e enriquecida pelos estágios anteriores para fornecer ainda mais detalhes. O rastreamento distribuído faz o uso mais intensivo do enriquecimento de telemetria nessa fase porque se concentra no uso de identificadores de correlação para vincular a telemetria. Dependendo dos detalhes técnicos, a vinculação da telemetria pode ocorrer no estágio ETC como parte do armazenamento, mas é o estágio de Apresentação que obtém essas correlações e as exibe para melhorar a tomada de decisões e a solução de problemas.

Até aqui, vimos que sistemas telemétricos são os sistemas que manipulam, transformam, armazenam e apresentam dados telemétricos. Ou seja, sua principal função é agregar e transformar os dados telemétricos das diversas fontes e apresentá-los para a tomada de decisões. Os quatro principais estilos de telemetria em uso hoje são: Registro Centralizado (*logging*), Métricas, Rastreamento Distribuído e Gerenciamento de Eventos de Segurança da Informação (GESI).

Cada um desses estilos segue a mesma arquitetura geral. No primeiro estágio, Recepção, todos os dados telemétricos auferidos são agregados e

preparados para uso dentro do seu pipeline de ETC de dados. Os sistemas que recebem as emissões do estágio anterior processam e transformam a telemetria para prepará-la para o armazenamento. Finalmente, no estágio Apresentação, os dados armazenados são analisados e visualizados, subsidiando decisões informadas para os stakeholders. Ao longo do pipeline, a telemetria é marcada com detalhes relacionados ao contexto e enriquecida para extrair detalhes codificados.

Com base nos conceitos e taxonomia vistos, vamos nos debruçar a seguir, sobre *loggers*. Os *loggers* agem principalmente como fontes de dados telemétricos (i.e., entradas para nossos sistemas de telemetria) e podem ser compreendidos como parte da primeira etapa do sistema, Recepção. Partindo de exemplos reais vamos ilustrar algumas implementações e como adaptá-las a IA. Contudo, trata-se apenas de ilustrações e o leitor não deve se limitar a elas para conformar seu próprio sistema telemétrico.

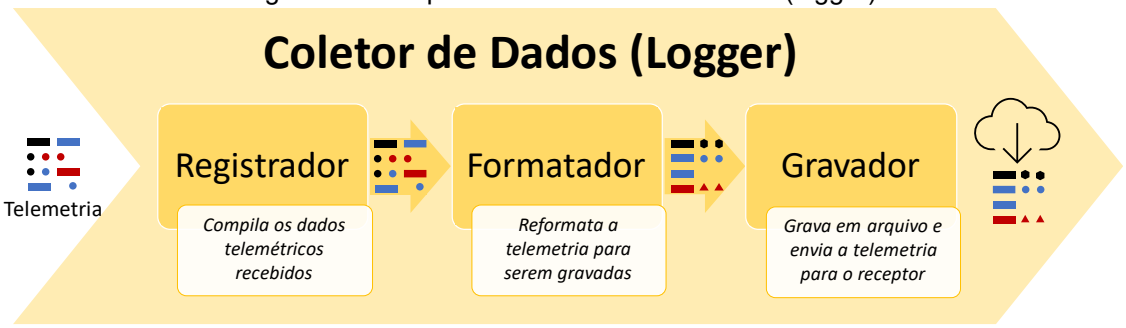
6.3.1.1.2 Logger estruturados

Os *loggers* (registradores ou coletor de dados) podem ser classificados em expedidores puros, *loggers* puros ou *loggers*/expedidores. Os expedidores puros são aqueles que despacham telemetria para um sistema diferente sem guardar cópia (i.e., funcionam como transmissores de dados agregados).

Os *loggers* puros fornecem telemetria para o mesmo sistema em que o código está sendo executado (e.g., a caixa preta dos aviões, *debugger* de uma IDE ou logs do Windows). Coletores de dados que despacham telemetria para um sistema diferente são *loggers*/expedidores (e.g., rastreador GPS do seu carro ou os dados de diagnóstico do Windows). Geralmente, os *loggers*/expedidores costumam guardar uma cópia local (i.e., arquivo de log) das telemetrias transmitidas mais recentemente.

Independente da classificação, os *loggers* seguem a arquitetura apresentada na Figura 6 – Componentes do coletor de dados (*logger*). Costumam diferir apenas na última fase, Gravador.

Figura 6 – Componentes do coletor de dados (logger)



Fonte: Riedesel (2021)

A figura acima, mostra os componentes de um *logger* genérico e como ele se relaciona com os conceitos de expedidor, registrador e registrador/expedidor. Os *loggers* estruturados possuem três componentes:

Registrador - atua como o arrolador da telemetria recebida pelas diversas fontes.

Formatador - edita a telemetria recebida no padrão adequado.

Gravador - grava localmente (*logger*) e envia a telemetria formatada para a próxima etapa (*logger*/expedidor).

Os registradores costumam ser encarados como fontes primárias de informações para os sistemas telemétricos. Contudo, na prática, são fontes secundárias, uma vez que apenas compilam telemetria recebida pelas reais fontes primárias. Essa ressalva é importante pois estes dados podem ser perdidos ou corrompidos durante o processo. A seguir, vamos ilustrar algumas implementações reais de *loggers* e como seu uso é vital, i.e., desde dirimir dúvidas até meio de prova.

Ilustração 1

Em 23 de março de 2018, uma falha no piloto automático da Tesla contribuiu para a morte de Walter Huang em Mountain View, Califórnia (NATIONAL TRANSPORTATION SAFETY BOARD OFFICE OF HIGHWAY SAFETY, 2019). Quando o Tesla Modelo X P100D (2017) de Huang se aproximou de uma saída à esquerda na US Highway 101, o software aparentemente confundiu as faixas na pista (NATIONAL TRANSPORTATION SAFETY BOARD OFFICE OF HIGHWAY SAFETY, 2020). O carro dirigiu-se para a esquerda, colocando-se entre os dois sentidos. Segundos depois, colidiu com uma divisória de concreto a 70 milhas por hora. Huang foi levado ao hospital, mas morreu logo depois (NATIONAL TRANSPORTATION SAFETY BOARD OFFICE OF HIGHWAY SAFETY, 2020).

Fonte: National Transportation Safety Board (2018)

Ilustração 2

Em 01 de março de 2019, o acidente de Delray ocorreu quando um veículo Tesla Model 3 (2018), que viajava para o sul na State Highway 441, atingiu um caminhão a leste que havia atravessado em frente ao veículo. No momento do acidente o sistema de "piloto automático" do Tesla estava ativo. O motorista, o único ocupante do carro, foi ferido fatalmente, o motorista do caminhão não ficou ferido (NATIONAL TRANSPORTATION SAFETY BOARD OFFICE OF HIGHWAY SAFETY, 2020).

Fonte: National Transportation Safety Board (2019)

Mas o que estes, e vários outros, acidentes envolvendo carros autônomos tem em comum? Fora envolverem modelos da Tesla, ambos possuíam módulos cuja função primária é prover registros para manutenção, reparo ou requerimentos operacionais específicos. Curiosamente, em todos os casos o piloto automático foi o principal suspeito pela morte dos passageiros. Mas, como é possível determinar a culpabilidade do piloto automático, geralmente um conjunto de inteligências artificiais com funções específicas?

Para fins de ilustração, no caso de Mountain View, o National Transportation Safety Board Office of Highway Safety (NTSB), trabalhando com a California Highway Patrol, recuperou dados do restraint control module (RCM); da media control unit (MCU); e tentaram extrair os dados do módulo Autopilot HW 2.5 electronic control unit (ECU) (NATIONAL TRANSPORTATION SAFETY BOARD OFFICE OF HIGHWAY SAFETY, 2020).

O RCM faz parte do sistema de restrição suplementar do Tesla. O módulo foi capaz de gravar dados quando acionado por um evento de acidente ou quase acidente, i.e., como um evento de acionamento do airbag. O *Event Data Recorder* (EDR) faz parte do RCM do veículo.

O Tesla Model X armazena dados, não georreferenciados, em memória não volátil usando um cartão SD removível, instalado no MCU do veículo. O cartão SD é grande o suficiente para manter um registro de todos os dados armazenados por mais de um ano. O cartão SD é, portanto, a principal base de dados do sistema de log do Tesla.

As informações gerais do veículo são gravadas continuamente no cartão SD (MCU) e registrados nos logs do veículo. Alguns dos parâmetros arrolados incluem ângulo de direção, posição do pedal do acelerador, aplicação do pedal do freio pelo motorista, velocidade do veículo, atributos do piloto automático (i.e., entradas da IA), acelerações longitudinal e lateral e distância do veículo da frente. Alguns desses parâmetros são gravados a uma taxa de 1Hz. Outros parâmetros são registrados apenas quando ocorre uma mudança de estado. Todos os parâmetros são marcados com a hora de chegada ao MCU usando um relógio derivado do GPS.

Parte dos dados contidos no cartão SD são sincronizados com os servidores da Tesla usando uma conexão de rede virtual privada (VPN), estabelecida via Wi-Fi ou usando os recursos de dados via celular no veículo. Notadamente, todos os dados armazenados só são extraídos por meio de equipamento de manutenção, conectado diretamente ao veículo ou removendo-se e acessando diretamente o Cartão SD, interno ao MCU montado no painel do carro.

Podemos inferir que o EDR da Tesla se trata dum *logger*/expedidor especializado desenvolvido para atender a demandas específicas das equipes de segurança, conformidade, desenvolvimento etc. que recolhe principalmente métricas enriquecidas com outros parâmetros, quando da mudança de estado (para exprimir o contexto).

Além dos dados mencionados acima, o veículo suporta o upload de dados geográficos de localização, anonimizados, para a Tesla, i.e., no intuito de melhorar a qualidade de mapas e do aprimoramento do piloto automático. Esses dados telemétricos não são armazenados a bordo do veículo.

O Autopilot HW 2.5 da Tesla, suporta 8 câmeras, 12 sensores ultrassônicos de longo alcance e um radar frontal. Um módulo processador (ECU) totalmente novo foi criado pela Tesla, com base no sistema Drive PX2 da Nvidia. Esta versão também possibilitou dois recursos que não são do piloto automático - câmera de painel e modo sentinela, com vídeo salvo localmente. Esta ECU, *logger* puro, está localizada abaixo do porta-luvas.

Se ao chegar aqui você notou uma leve, se não grande, semelhança entre esses componentes e a caixa-preta dos aviões, saiba que não é mera coincidência. Ambos são dispositivos usados para registrar parâmetros específicos de desempenho.

Flight Data Recorders (FDR), gravadores de dados de voo – conhecidos popularmente como caixa preta e às vezes conhecidos como gravadores de dados de acidentes – são dispositivos usados para registrar dados sobre a aeronave, que podem ser usados para investigação de acidentes e para analisar questões de segurança aérea, degradação de material e desempenho do motor.

Os FDR são regulamentados pela *International Civil Aviation Organization* (ICAO), i.e., pelas normas internacionalmente reconhecidas e pelas práticas recomendadas contidas no Anexo 6 da ICAO (INTERNATIONAL CIVIL AVIATION ORGANIZATION, 2010). Essas especificações podem ser encontradas na Organização Europeia para Equipamento de Aviação Civil (Eurocae) ED112 (EUROPEAN ORGANIZATION FOR CIVIL AVIATION EQUIPMENT, 2013). Nos Estados Unidos, a *Federal Aviation Administration* (FAA) cita os requisitos de projeto em seu *technical standard order* TSO-C124b (FEDERAL AVIATION ADMINISTRATION, 2019), que é baseado em documentos da Eurocae.

Os FDR registram parâmetros significativos de voo, incluindo posições de controle e atuador, desempenho do motor e hora do dia. Existem 88 parâmetros obrigatórios de acordo com os regulamentos atuais dos EUA (INTERNATIONAL CIVIL AVIATION ORGANIZATION, 2010); no entanto, alguns sistemas monitoram muito mais variáveis. Cada parâmetro adicional que o FDR registra, fornece aos investigadores mais uma pista sobre a causa de um acidente.

Os FDR são um exemplo clássico de métricas em *logger* puro. Eles se concentram em parâmetros (contadores, cronômetros, taxas etc.), em vez de texto, para descrever o que está acontecendo, o que lhes permite prazos muito maiores de dados armazenados.

Geralmente, um FDR é montado na seção traseira de uma aeronave – onde é mais provável que sobreviva a um acidente grave – e outra unidade abaixo do cockpit. Os exemplares modernos são envoltos duas vezes em aço inoxidável ou titânio, forte e resistente à corrosão, capaz de suportar uma aceleração de 3.400 Gs (33 km/s ao quadrado) por 6,5 milissegundos, e no interior, existe isolamento de alta temperatura. As unidades são projetadas para emitir um sinal de localizador por até 30 dias e podem operar a uma profundidade de 6.000 metros (20.000 pés) (INTERNATIONAL CIVIL AVIATION ORGANIZATION, 2010).

O *Cockpit Voice Recorder* (CVR), tem por finalidade gravar a conversa de cada membro da tripulação, transmitida e recebida, com instalações terrestres e no sistema de intercomunicação do avião e reter os últimos 30 minutos da conversa da tripulação. Além disso contém medidas para parar o gravador em caso de falha, para que os últimos 30 minutos de conversa não sejam apagados ou substituídos e indicativos para informar a tripulação quando estiver operando corretamente (INTERNATIONAL CIVIL AVIATION ORGANIZATION, 2010).

Os CVR são um exemplo clássico de registro centralizado em *logger* puro. As peculiaridades aqui são duas. A primeira é que eles se concentram em voz, em vez de texto, para descrever o que está acontecendo, o que lhes permite prazos muito curtos de dados armazenados. A segunda é quanto a possibilidade de apagar os logs, só efetivando quando há o correto encerramento das atividades da aeronave por pessoa autorizada.

ATC Communication Recordings são as gravações originadas na torre de comando. Esses registros, mantidos por 30 dias, englobam as comunicações entre controladores e pilotos – i.e., *Controller Pilot Data Link* (CPDL) – e as comunicações entre controladores. No caso de um acidente ou incidente, a gravação original do evento pode ser requisitada pelos investigadores; caso contrário, o meio de gravação será reutilizado e as informações perdidas.

A importância dos dados gravados da ATC será determinada pelas circunstâncias que envolvem um acidente ou incidente, e.g., correlação de eventos entre às gravações ATC, FDR e CVR pode fornecer referência de tempo e espaço muito precisas. Além da referência de tempo, as informações gravadas da ATC também fornecem referências da pista, particularmente úteis na avaliação da posição relativa do avião quando várias aeronaves estão envolvidas.

Os FDR do futuro serão obrigados a gravar mais do que os parâmetros tradicionais de voz e dados. Os recentes avanços na tecnologia de vídeo tornaram a gravação de vídeo, tanto da cabine quando de partes da aeronave, uma possibilidade. Além disso, a CPDL atingirá um nível que exigirá a gravação dessas mensagens pelo gravador de voo. Isso mostra uma tendencia de evolu-

ção do estilo de registro centralizado para sistemas de gerenciamento de eventos telemétricos sofisticados.

Assim como nos Teslas, temos uma miríade de FDR mandatórios e específicos, com suas respectivas funções; que auxiliam os investigadores e fornecem mais pistas sobre as causas de um acidente. Igualmente ao piloto automático do Tesla, agente da celeuma hodierna, seu símile aeronáutico nos ilustra o que esperar para a contenda da responsabilização civil.

O principal obstáculo social nos primórdios da aviação era a crença generalizada na incapacidade humana de controlar uma aeronave. Segundo o jurista Henry Grady Jr. Gatlin,

---

"A aviação em seu estágio atual de desenvolvimento é extremamente perigosa, porque mesmo o melhor avião, construído e mantido, é tão incapaz de controle completo que voar cria um risco de que o avião, embora cuidadosamente construído, mantido e operado, possa colidir ocasionando ferimentos às pessoas, estruturas e bens móveis em terra ..." (VOLD, 1953).

Esse problema começou a ser sanado com o advento do piloto automático.

O primeiro piloto automático é creditado a Lawrence Sperry (SCHECK, 2004). O sistema de Sperry era um piloto automático operando em apenas dois dos três eixos de controle do avião, o que permitia à aeronave voar em linha reta e nivelada em um curso, sem a necessidade da atenção do piloto. Seu funcionamento baseava-se um mecanismo interligado, que media a posição da aeronave, a um motor que movia os cabos conectados aos controles da aeronave. O primeiro voo realizado com sucesso ocorreu em Nova York em 1913 e foi publicamente demonstrado em 18 junho de 1914 no *Concours de la Sécurité en Aéroplane* em Paris (SCHECK, 2004).

O piloto automático da Sperry e tecnologias semelhantes, agraciaram os dispositivos como revolucionários para a segurança aérea (NEW YORK TIMES, 1916). A ênfase na segurança era compreensível, uma vez que voar era uma atividade de extremo perigo naquele momento e geralmente visto como um hobby para ricos e insanos (RICH, 1993).

33 anos depois, em 1947, uma aeronave militar americana realizou o primeiro voo totalmente automático – sem intervenção humana, da decolagem à aterrisagem – através do Oceano Atlântico, voando de Newfoundland para a Inglaterra.

O voo se deu sobre o controle do piloto automático ("o Cérebro") – um computador programado por cartões perfurados – e contava com uma equipe de bordo para verificar o percurso e observar os instrumentos. A maioria deles tinha pouco ou nada a fazer durante o trajeto (TIMES, 1947). Qualquer semelhança com a metodologia empregada pela Waymo[53] para testar seus carros autônomos não é mera coincidência (WOOLLASTON, 2016).

À medida que as habilidades dos pilotos automáticos e de outros sistemas automatizados de aviação se tornaram mais sofisticados e amplamente utilizados, a automação passou a ser vista como essencial para a segurança e o sucesso da aviação. Consequentemente, foi a própria indústria aeronáutica, e grupos associados, que reivindicaram a certificação federal e os regulamentos de segurança.

Na prática, a responsabilidade de pilotar a aeronave foi distribuída entre a tripulação e a automação, com a última em posição de destaque. Sob esse enfoque, poderíamos esperar um novo regime de prestação de contas e responsabilização, e que os responsáveis pelo desenvolvimento, manufatura, instalação e manutenção dessas automações, assumissem maiores responsabilidades sobre falhas nos voos (COOLING e HERBERS, 1983); no entanto, este não foi o caso.

---

[53] A Waymo é uma subsidiária da Alphabet Inc. que desenvolve tecnologia para carros autônomos.

Os primeiros casos envolvendo piloto automático, apresentavam as companhias aéreas como réus, mesmo quando se suspeitava do piloto automático como causa do acidente, e exemplifica uma classe de processos no qual a companhia aérea é considerada estritamente responsável. Às vezes, os fabricantes aparecem como réus no litígio, quando há acusação de projeto negligente, ou falha, em avisar a tripulação de risco iminente. Todavia, a classe prevalente envolve a utilização do piloto automático pelo comandante. Essencialmente, os casos são em torno da falha ou uso (indevido ou não) do piloto automático pelo comandante.

À medida que os pilotos automáticos se tornaram centrais nos voos comerciais, os tribunais incorporaram as novas tecnologias aos regimes de responsabilização existentes, sem maiores contendas. Embora esses primeiros casos sejam indelineáveis, os regulamentos deixam claro que a responsabilidade e o risco são do operador, i.e., comandante, e não do sistema automatizado.

No corpo dos requerimentos para certificação da aviação civil, há duas seções de fundamental importância para entender como os tribunais trataram as questões de responsabilização relacionadas à pilotos automáticos.

A primeira, 14 CFR 23.1329, lista os requisitos que o design de qualquer piloto automático deve atender: Ser rápido e positivamente desativado pelos pilotos para evitar que interfira no controle do avião; e ser suficientemente subjugável por um piloto, para deixá-lo controlar o avião (FEDERAL AVIATION ADMINISTRATION, 2002).

A segunda, 14 CFR 91.3, diz que "o piloto no comando de uma aeronave é diretamente responsável e é a autoridade final quanto à operação dessa aeronave" (FEDERAL AVIATION ADMINISTRATION, 2006). O piloto (e, por extensão, na maioria dos casos, a companhia aérea) é responsável pela operação do avião e pelo uso (ou não), operação (correta ou não) e cumprimento (ou não) do piloto automático.

Dessa forma, nos vemos diante de um cenário em que a automação é vista como segura (e superior na maioria dos casos), a menos que algo dê er-

rado, quando os humanos (aí considerados mais seguros e superiores) devem corrigir e assumir a situação.

Infelizmente, pôr as pessoas nessa conjuntura – i.e., onde devem resolver uma situação emergencial no último instante – não é o tipo de tarefa que elas executem bem, ou mesmo possam fazer, e.g., tempo insuficiente para reação.

Um exemplo onde realmente houve falha humana em detectar e interpretar o piloto automático, i.e., ter seguido as instruções do piloto automático a tempo teria salvado os passageiros, é o do Voo 3407 da Colgan Air. Onde a National Transportation Safety Board (NTSB) concluiu que a resposta inadequada do comandante à ativação do *stick shaker*[54] levou a uma perda de sustentação da qual o avião não se recuperou (NTBS, 2009).

Caso semelhante, parece ser o do acidente envolvendo o carro autônomo da Uber, que em março de 2018 vitimou a pedestre Elaine Herzberg em Tempe, Arizona. Diferentemente do caso anterior, o NTSB decidiu que a culpa no caso é da Uber, por não avaliar corretamente os riscos de segurança (visto que não possuía um departamento específico para avaliação e mitigação de riscos na época do acidente); da motorista de segurança, que estava mexendo em seu celular pessoal até um segundo antes da colisão; do governo do Arizona, por políticas insuficientes para a regulamentação dos carros autônomos em vias públicas; e da própria vítima, por estar sob efeito de metanfetamina e ter atravessado a rua fora da faixa de pedestre (o que contribuiu para o acidente) (HAWKINS, 2019).

Esses dois casos nos alertam para duas necessidades. A primeira é a de previsões legais para esses novos casos de responsabilização (o que complementa, porém, foge do escopo desse trabalho). A segunda é a da importância dos módulos cuja função primária é prover registros para manutenção, reparo ou requerimentos operacionais específicos.

---

[54] Um dispositivo de alerta de perda de sustentação que, além de avisar o piloto, também pode recuperar automaticamente a aeronave da perda de sustentação (KUMAR, 2005).

6.3.1.1.2.1 Explainable Accountable Unit (XAU)

A *Explainable Accountable Unit* (XAU), unidade auditora explicável (em português), consiste num coletor de dados desenvolvido para atender as demandas específicas dos diversos stakeholders envolvidos no ciclo de vida da IA, i.e., deve sempre registrar dados sobre a IA que posam ser analisados e usitados para investigação de ocorrências, prestação de contas, depuração, imputabilidade etc.

Essa abordagem já se consagrou na literatura aeronáutica, automotora, náutica e, como vimos em vários casos anteriormente, se mostrou útil para elucidar as verdadeiras causas e a apontar responsáveis em diversos acidentes envolvendo carros autônomos. Por isso, parece-nos viável estender o conceito às demais aplicações da IA observando-se sempre as idiossincrasias de cada contexto.

Para investigar minuciosamente os fatores associados a um incidente, usamos algumas técnicas de investigação, como: preservação de evidências, depoimentos e entrevista com testemunhas. A contraparte IA para as duas primeiras técnicas será abordada pela *Explainable Accountable Unit* (XAU), e a última pelas técnicas XAI (próximo capítulo).

A preservação de evidências visa preservar, coletar e documentar com sucesso as evidências que podem contribuir para a compreensão do acidente. A eficácia de uma investigação depende da preservação imediata do local, das evidências físicas, humanas, cibernéticas e documentais relacionadas ao incidente, bem como de sua segurança e custódia, para evitar adulterações ou perdas e estabelecer precisão e validade.

O depoimento de uma testemunha envolvida em um acidente geralmente contém informações sobre o que o indivíduo viu, mas podem conter outras informações pertinentes não diretamente ligadas ao que aconteceu. A declaração é assinada pela testemunha, garantindo a sua autenticidade. Esse registro também é usado para determinar a hora, o local e a sequência de eventos, cruciais para determinar a causa raiz.

Habilitar a gravação automática de um evento da IA é o que temos de mais próximo das declarações de recordação de testemunhas. Além disso, as telemetrias coletam, documentam e preservam evidências que podem contribuir para a compreensão do acidente.

Pode-se perguntar se simplesmente ter um registro é suficiente para estabelecer uma explicação. Tal informação não é suficiente por si só se estamos buscando uma explicação completa e detalhada, por exemplo, a causa raiz, mas é um pré-requisito essencial para estabelecer se um risco tecnológico se materializou e pode servir como base para nossas técnicas de XAI.

Equipar um sistema de IA com os meios para registrar informações operacionais atinge um certo grau de explicabilidade. A principal ideia por trás do log é a capacidade de comparar um conjunto cronológico de entradas e saídas para fornecer uma interpretação do que aconteceu, ou seja, a cadeia de eventos que levaram a um fato.

A XAU é responsável por coletar, armazenar, tratar e comunicar informações telemétricas (entradas, saídas, estados etc.) essenciais para recriar as condições de operação de um sistema de IA em um determinado momento. Sua operação primária é como a do FDR, que preserva o histórico de voo recente, registrando dezenas de parâmetros, coletados várias vezes por segundo. Outras informações sobre o funcionamento do sistema, não vinculadas à IA, devem ser consideradas ao descrever o contexto (natureza, magnitude, localização e tempo), por exemplo, imagens de áudio e vídeo, coordenadas de GPS ou condições climáticas.

A implementação da XAU deve ser conduzida de forma que nenhuma parte interessada possa manipular os dados, enquanto todas as partes interessadas mantêm o acesso conforme necessário. Além disso, a elaboração deve considerar quaisquer implicações adversas para os direitos de terceiros e ser conduzida de acordo com as leis aplicáveis, e.g., Regulamento Geral de Proteção de Dados (privacidade), Agência Europeia para a Segurança da Aviação (domínio específico) e segredos comerciais.

Questões técnicas como: custo de armazenamento e transmissão, viabilidade técnica e meios alternativos de coleta de informações, devem ser consideradas ao solicitar, especificar e projetar a XAU. As agências reguladoras (específicas para as aplicações apropriadas, por exemplo, dispositivos médicos, regulamentação de máquinas, aviação civil, veículos automotores etc.), por terem o conhecimento adequado e equipe técnica treinada, devem liderar a padronização e outras atribuições da XAU para aplicações específicas.

Devemos também considerar o momento da coleta, i.e., ex-ante e ex-post. Todos os casos que abordamos até agora visam uma análise ex-post. No entanto, se quisermos implementar um ciclo virtuoso de *accountability*, devemos empregar a XAU durante todo o ciclo de vida da IA. Para nossa satisfação, os diversos estilos cobertos, vide item 6.3.1.1 Sistemas de telemetria, são capazes de obter, manipular, transformar, armazenar e apresentar dados telemétricos das diversas fases e tarefas executadas pelo processo discutido, vide item 6.2.1 CRoss-Industry Standard Process Model (CRISP-DM).

Neste capítulo aprendemos quando, onde e como coletar informações sobre a IA. Começamos por apresentar os conceitos de ex-ante e ex-post. Em seguida, analisamos a principal metodologia de data mining empregada na atualidade, CRoss-Industry Standard Process Model (CRISP-DM). Por fim, investigamos as principais arquiteturas de telemetria de software que podemos empregar para coletar, manipular, transformar, armazenar e apresentar dados obtidos ao longo do ciclo de vida da IA culminando numa abordagem para a IA, a Explainable Accountable Unit (XAU).

Saneadas as questões de como operacionalizar a coleta de dados, no capítulo seguinte, nos debruçaremos sobre quais tipos de informações serão coletados e quais divulgações serão feitas. Isso impactará não só a especificação da XAU, mas o sistema telemétrico e o processo de produção da IA.

7 QUAIS TIPOS DE INFORMAÇÕES SERÃO COLETADOS E QUAIS DIVULGAÇÕES SERÃO FEITAS?

Enquanto demandados, devemos sempre ser honestos sobre a exequibilidade do pactuado. Devemos manter um diálogo franco e sólido com os solicitantes, apresentando contraofertas viáveis (e.g., redução de escopo, aumento de prazo, compartilhamento ou redistribuição de tarefas, condicionamento de resultados etc.) até encontrar um denominador aceitável para ambos. Ao demandante, caberá reagir ao que foi proposto, apresentar contraofertas e garantir que está recebendo uma promessa de qualidade (i.e., factível e que atenda ao máximo suas expectativas).

Tudo correndo bem, demos nossa palavra ao solicitante, que sente confiança no que foi pactuado. Isso pode variar de um “Combinado!” ou um aperto de mão entre as partes, até declarações públicas (e.g., carta de intenções, coletiva de imprensa, contrato etc.).

O pior que pode acontecer é quando o trato é estabelecido de forma unilateral pelo demandante, que deposita uma série de expectativas irreais sobre o que será entregue, ou alijando stakeholders importantes da parte demandada (e.g., apenas os diretores participam das tratativas e excluem as partes que realmente executaram o trato). Quando isso acontece, ambos os contraentes se sentirão lesados e, geralmente, jogarão a culpa do eventual fracasso um no outro.

Não obstante, o fato de uma promissão ser feita não garante sua realização. As partes são responsáveis por garantir que o plácito seja cumprido. Devem atinar para externalidades (e.g., circunstâncias, desafios, prioridades etc.), limitar danos e manterem-se focadas em lograrem os resultados pactuados.

Como versado ao longo dessa obra, ao instado, caberá realizar certas ações, ou abster-se de, de acordo com as expectativas pactuadas e fornecer

uma avaliação dessas ações para os pleiteantes. O requisitante, deve acompanhar o progresso e manter-se estoico quanto ao convencionado.

À vista disso, nos ocuparemos de responder ao longo desse capítulo a (QN3) “Quais tipos de informações serão coletados e quais divulgações serão feitas?”. Dada a correlação entre o que se coleta e o que se divulga, por uma questão didática, abordamos primeiramente quais divulgações serão feitas (item 7.1 Quais divulgações serão feitas) para em seguida apresentarmos e diferenciarmos o escopo: os cômputos internos (7.1.1 Cômputos internos), i.e., aqueles que fazemos para diretores, programadores etc., e externos (7.1.2 Cômputos externos), i.e., aqueles que fazemos para investidores, consumidores, judiciário etc.

Quanto a segunda parte, “Quais tipos de informações serão coletados”, desenvolvemos os recursos manuseados para confecção dos proclames. No item 7.2.1 O que medir expomos o que medir, no tópico 7.2.2 Onde medir retratamos onde medir, i.e., cadeia de suprimentos (7.2.2.1 Cadeia de suprimentos) e Análise do Ciclo de Vida (7.2.2.2 Análise do Ciclo de Vida (ACV)), e o ponto 7.2.3 Como medir se ocupa de como medir.

Coletar e relatar alguns desses recursos de natureza potencialmente relevante e representativamente fiel pode ser caro, consequentemente, precisamos considerar os possíveis custos e benefícios associados à produção das informações. Os benefícios associados à divulgação devem exceder os custos de disponibilização dessa informação. Dessarte, custos são abordados no item 7.2.3.1 Custo e sua aplicação ao ciclo de vida é abordado em 7.2.3.2 Análise de Custo do Ciclo de Vida (ACCV).

Definidos escopo e recursos, reservamos o terceiro item 7.3 XAI, para expor arcabouço equivalente aplicado a IA. Nele expomos o conceito de explicabilidade 7.3.1 Explicabilidade, uma taxonomia dos seus métodos 7.3.1.1 Taxonomia dos Métodos de Explicabilidade, algumas de suas principais técnicas 7.3.1.2 Técnicas de Explicabilidade e um subconjunto temático, sobre imparcialidade, no item 7.3.2 Imparcialidade.

7.1 QUAIS DIVULGAÇÕES SERÃO FEITAS

No presente, existe uma ampla variedade de organizações envolvidas em diferentes atividades, com variegados grupos de interesse e com obrigações e responsabilidades percebidas desconformes, empenhando recursos e concebendo algoritmos sui generis. Como tal, não devemos perceber a divulgação como uma prática uniforme para todos. A seguir, analisaremos as duas modalidades de cômputos produzidos para os stakeholders: cômputos internos (ou contabilidade gerencial) e cômputos externos (e.g., relatórios financeiros, relatórios de impacto etc.).

7.1.1 Cômputos internos

A contabilidade gerencial é um termo amplo que se refere às informações produzidas para a tomada de decisão interna pelos gestores (i.e., informações necessárias para diversas decisões referentes a aspectos associados ao planejamento, monitoramento e controle, incluindo aspectos de desempenho financeiro, social e ambiental). Ela é produzida internamente em vários formatos e conforme solicitada pelas várias partes que se valem dessas informações.

Como essas informações são adaptadas para atender às necessidades específicas dos diferentes gestores, a contabilidade gerencial realizada dentro das organizações matiza (i.e., os cômputos produzidos pelas diversas organizações variam em qualidade, conteúdo, formato, apresentação etc.). Contudo, embora diferentes na forma, os relatórios gerenciais são funcionalmente equivalentes. Isto é, se os gerentes quiserem ter sucesso, eles devem realizar uma análise cuidadosa de todas as entradas e saídas de seu processo de produção e se concentrarem em reduzir os aspectos negativos de suas operações.

Essa contabilidade gerencial deriva, sobretudo, de mensurações e, portanto, lança mão de variegada estrutura de diagnóstico para aferir e relatar os

diferentes aspectos analisados. Assim, gera-se um pool de informações para os gerentes que serão de interesse (ou não) para uma variedade de stakeholders, dentro e fora da organização. Embora muitos stakeholders estejam interessados nas informações usadas internamente, por motivos de sigilo ou vantagem competitiva, essas informações nem sempre podem ser compartilhadas ou divulgadas.

No entanto, é preciso lembrar que o público-alvo das contas gerenciais são, em última análise, os gestores. Se eles optam por divulgar informações de seus cômputos gerencias para partes interessadas externas, isso é motivado pela crença de que são responsáveis perante esses stakeholders em relação a aspectos específicos do desempenho ou por acreditarem que é do interesse organizacional divulgar essas informações. Além disso, conforme indicado anteriormente, considerações sobre vantagem competitiva também influenciarão a divulgação pública de certas informações.

7.1.2 Cômputos externos

Com os interesses financeiros dos investidores, credores e outros interessados em mente, a contabilidade financeira é altamente regulamentada com relação aos procedimentos a serem usados para gerar demonstrações financeiras de propósito geral (i.e., relatórios financeiros externos). Isso pode ser contrastado com a contabilidade gerencial, que não é regulamentada porque os gerentes são mais do que capazes de determinar quais elementos (i.e., contabilidade gerencial) precisam ou desejam para gerenciar a organização.

Ao contrário dos relatórios financeiros, ainda há pouca regulamentação que rege os relatórios externos de desempenho da IA e, portanto, grande parte da análise realizada fica a critério da organização. Isso, por sua vez, pode ser influenciado por valores pessoais, pressão das partes interessadas, escassez e / ou natureza dos dados empregados, formação profissional e educacional, e assim por diante.

No que concerne a imputabilidade, fatores como receita bruta anual, número de consumidores, natureza jurídica ou atividade econômica determinam sua aplicação. Como já explanado, tudo isso afetará a necessidade bem como as contas apresentadas. Por exemplo, o *Algorithmic Accountability Act of 2019* (AAA19), pretende-se aplicar a qualquer pessoa, parceria ou corporação sobre a qual a Federal Trade Commission (FTC) tenha jurisdição que:

- Tenha mais de US$ 50.000.000,00 de receitas brutas anuais médias para o período de três anos tributáveis anterior ao ano fiscal mais recente;
- Possua ou controle informações pessoais de mais de 1.000.000 de consumidores ou 1.000.000 de dispositivos de consumidores;
- Possua, seja operado ou controlado substancialmente por uma pessoa, parceria ou corporação; ou
- Seja um corretor de dados ou outra entidade comercial que, como parte substancial de seus negócios, coleta, reúne ou mantém informações pessoais sobre um indivíduo que não é um cliente ou funcionário dessa entidade para vender ou negociar as informações ou fornecer acesso a terceiros (UNITED STATES, 2019).

Apesar de não restrita a IA, a LGPD e o RGPD acabam por balizar certos papeis (i.e., funções na tratativa de dados) nos quais eventualmente possamos nos enquadrar. Apesar de alguma divergência nos critérios de enquadramento, os direitos e deveres coabitam os mesmos princípios.

Para a Lei Geral de Proteção de Dados,

Controlador é pessoa natural ou jurídica, de direito público ou privado, a quem competem as decisões referentes ao tratamento de dados pessoais;

Operador é pessoa natural ou jurídica, de direito público ou privado, que realiza o tratamento de dados pessoais em nome do controlador;

Encarregado é a pessoa indicada pelo controlador e operador para atuar como canal de comunicação entre o controlador, os titulares dos dados e a Autoridade Nacional de Proteção de Dados (ANPD);

Agentes de tratamento são o controlador e o operador; e

Órgão de pesquisa é o órgão ou entidade da administração pública direta ou indireta ou pessoa jurídica de direito privado sem fins lucrativos legalmente constituída sob as leis brasileiras, com sede e foro no País, que inclua em sua missão institucional ou em seu objetivo social ou estatutário a pesquisa básica ou aplicada de caráter histórico, científico, tecnológico ou estatístico (BRASIL, 2019).

Para o Regulamento Geral de Proteção de Dados,

Responsável pelo tratamento, é a pessoa singular ou coletiva, a autoridade pública, a agência ou outro organismo que, individualmente ou em conjunto com outras, determina as finalidades e os meios de tratamento de dados pessoais; sempre que as finalidades e os meios desse tratamento sejam determinados pelo direito da União ou de um Estado-Membro, o responsável pelo tratamento ou os critérios específicos aplicáveis à sua nomeação podem ser previstos pelo direito da União ou de um Estado-Membro;

Subcontratante, é uma pessoa singular ou coletiva, a autoridade pública, agência ou outro organismo que trate os dados pessoais por conta do responsável pelo tratamento destes;

Destinatário, é uma pessoa singular ou coletiva, a autoridade pública, agência ou outro organismo que recebem comunicações de dados pessoais, independentemente de se tratar ou não de um terceiro. Contudo, as autoridades públicas que possam receber dados pessoais no âmbito de inquéritos específicos nos termos do direito da União ou dos Estados-Membros não são consideradas destinatários; o tratamento desses dados por essas autoridades públicas deve cumprir as regras de proteção de dados aplicáveis em função das finalidades do tratamento;

Terceiro, é a pessoa singular ou coletiva, a autoridade pública, o serviço ou organismo que não seja o titular dos dados, o responsável pelo tratamento, o subcontratante e as pessoas que, sob a autoridade direta do responsável pelo tratamento ou do subcontratante, estão autorizadas a tratar os dados pessoais;

Representante, é uma pessoa singular ou coletiva estabelecida na União que, designada por escrito pelo responsável pelo tratamento ou subcontratante, nos termos do artigo 27º, representa o responsável pelo tratamento ou o subcontratante no que se refere às suas obrigações respetivas nos termos do presente regulamento;

Empresa, é uma pessoa singular ou coletiva que, independentemente da sua forma jurídica, exerce uma atividade económica, incluindo as sociedades ou associações que exercem regularmente uma atividade económica;

Grupo empresarial, é um grupo composto pela empresa que exerce o controle e pelas empresas controladas;

Serviços da sociedade da informação, é um serviço definido no artigo 1º, nº 1, alínea b), da Diretiva (UE) 2015/1535 do Parlamento Europeu e do Conselho (19);

Organização internacional, é uma organização e os organismos de direito internacional público por ela tutelados, ou outro organismo criado por um acordo celebrado entre dois ou mais países ou com base num acordo dessa natureza (EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2016).

Essas duas legislações costumam ser o foco da atenção dos principais desenvolvedores nacionais. Contudo, devido a transnacionalidade do alcance da IA no cenário atual, deve-se observar, especialmente quando ofertada como bem ou serviço, a legislação local.

Países como África do Sul (Protection of Personal Information Act - POPI), Argentina (Ley de Acceso a la Información Pública, Protección de Datos

Personales y el Registro Nacional - PDP), Canadá (The Personal Information Protection and Electronic Documents Act - PIPEDA), China (Cyber Security Law), Japão (Act on the Protection of Personal Information - APPI) etc., possuem legislação una com abrangência nacional. Ao passo que países como Austrália e Estados Unidos regulam a privacidade e a proteção de dados por meio de uma combinação de leis federais, estaduais e territoriais.

Exempli gratia, a lei de privacidade federal australiana (Privacy Act) aplica-se a entidades do setor privado, incluindo corporações, parcerias, fundos e associações não incorporadas, com um faturamento anual de pelo menos AU$ 3 milhões e todas as agências do Governo do Commonwealth e do Território da Capital Australiana (AUSTRALIAN GOVERNMENT, 2019). Já a maioria dos estados e territórios, exceto Austrália Ocidental e Austrália do Sul, têm sua própria legislação de proteção de dados aplicável a agências governamentais estaduais e empresas privadas que interagem com agências governamentais estaduais.

Mesmo que as referidas leis não tratem de inteligência artificial diretamente, elas legislam sobre seu principal insumo e tratativa cabendo, por tanto, conformidade. Portanto, incorremos numa série de responsabilizações e prestação de contas involuntárias, mas que devem, obrigatoriamente, constar em nossos cômputos externos para fins de conformidade.

A organização normalmente usará uma pluralidade de estruturas contábeis / descritivas para fornecer os vários cômputos que se relacionam aos heterogêneos aspectos da IA. Para a geração de seus relatórios, lançará mão de práticas acreditadas, ou mesmo demandadas, por seus stakeholders.

O AAA19 orienta a FTC a exigir das entidades – que usam, armazenam ou compartilham informações pessoais – a realização da avaliação de impacto do sistema de decisão automatizada e avaliação de impacto na proteção de dados (UNITED STATES, 2019).

Para a referida lei, a avaliação de impacto na proteção de dados, significa um estudo que avalia até que ponto um sistema de informação protege a privacidade e a segurança das informações pessoais que o sistema processa.

Já a avaliação de impacto do sistema de decisão automatizada consiste em um estudo que avalia o sistema de decisão automatizada e o processo de desenvolvimento do sistema, incluindo design e dados de treinamento do modelo, quanto a impactos na precisão, justiça, preconceito, discriminação, privacidade e segurança. Deve incluir, mas não se limita a:

> (k1) uma descrição detalhada do sistema, i.e., seu design, treinamento, dados e finalidade;
>
> (k2) uma avaliação dos benefícios e custos relativos ao sistema, à luz de sua finalidade, levando em consideração fatores relevantes, incluindo:
>
> (i) práticas de minimização de dados;
>
> (ii) a duração pela qual as informações pessoais e os resultados do sistema são armazenados;
>
> (iii) quais informações sobre o sistema estão disponíveis para os consumidores;
>
> (iv) até que ponto os consumidores têm acesso aos resultados do sistema e podem corrigir ou objetar seus resultados; e
>
> (v) os destinatários dos resultados do sistema;
>
> (k3) uma avaliação dos riscos impostos pelo sistema à privacidade ou segurança das informações pessoais dos consumidores e aos riscos que o sistema pode resultar ou contribuir para decisões imprecisas, injustas, tendenciosas ou discriminatórias que afetam os consumidores; e
>
> (k4) as medidas que a entidade coberta empregará para minimizar os riscos descritos no subparágrafo (k3), incluindo salvaguardas tecnológicas e físicas.

A exigência da realização, tanto para sistemas novos quanto existentes, das avaliações de impacto do sistema de decisão e na proteção de dados, será mandatória após 2 anos de promulgada a lei (UNITED STATES, 2019). O AAA19 enfatiza, mas não obriga, a importância desses estudos serem realizados em consulta com terceiros, incluindo auditores e especialistas em tecnolo-

gia independentes. Já os documentos, podem ser tornados público pela entidade a seu exclusivo critério (UNITED STATES, 2019).

Até aqui, vimos as duas classes de divulgações (internas e externas). Enquanto as divulgações externas são dependentes quase que unicamente dos desejos da diretoria as externas estão, cada vez mais, migrando para uma regulamentação forte e detalhada. Contudo, independentemente do seu escopo, resta a tarefa de coletar os dados que embasam essas divulgações, i.e., quais tipos de informações serão coletados?

7.2 QUAIS TIPOS DE INFORMAÇÕES SERÃO COLETADOS

Completa a tarefa de decidir quais divulgações serão feitas, é hora de se estabelecer o que medir, como medir e onde medir. Os tipos de informações a serem coletadas, analisadas e relatadas dependerão, em parte, de quais aspectos do desempenho serão acompanhados. Isso será influenciado por vários fatores, incluindo:

- As responsabilidades assumidas pelos encarregados;
- Os limites de reporte (influenciado pelas responsabilidades assumidas pelos gerentes);
- O tipo de produtos ou serviços oferecidos pela organização;
- Os tipos de recursos consumidos e a natureza dos resultados/impactos sendo criados e medidos;
- A localização das atividades da organização e os impactos das operações da organização nos diferentes meios;
- O nível de competição que a organização se encontra.

Se a coleta e análise das informações relevantes for dispendiosa e se o benefício para a tomada de decisão for percebido como mínimo, uma decisão racional seria não coletar as informações. Embora coletar poucas informações

seja ruim para a tomada de decisões, coletar e fornecer grandes quantidades de informações também pode prejudicar a tomada de decisões.

Portanto, é importante que as pessoas envolvidas na função “contábil” restrinjam as informações disponíveis àquelas que sejam relevantes e compreensíveis pelos interessados. Além de ser relevante, confiável e compreensível, as outras características qualitativas desejadas para essas informações são comparabilidade, verificabilidade e temporalidade. Os três principais critérios que devem ser atendidos pelos dados a serem apurados são:

(1) Possibilidade de coleta de forma confiável;

(2) Relevância para a organização;

(3) Controle de acesso.

7.2.1 O que medir

Assim como uma organização usará vários recursos como parte de suas operações e criará vários resultados (ou impactos), alguns que são pretendidos (como bens e serviços) e alguns adversos (e.g., resíduos, ou poluição, que podem criar custos externos), também faz a IA. Os resultados produzidos por uma IA podem incluir: detecção (e.g., câncer em pacientes ou furos em filmes plásticos), predição (e.g., tráfego ou sugestão de produtos), geração (e.g., música, vídeos, notícias ou quadro), estimativa (e.g., distâncias, falsificação ou confiabilidade de uma amostra), planejamento (e.g., rotas de entrega, otimização de recursos ou escalonamento) ou reconhecimento (e.g., facial, voz, leitura labial, contexto ou mesmo gerar legendas automaticamente).

Como vimos, a decisão sobre quais entradas e saídas medir e relatar realmente depende do propósito subjacente da IA e das metas das pessoas que a gerenciam. De uma perspectiva interna, os administradores desejam tomar decisões embasadas para atingir metas específicas e precisam de informações para fazer isso, i.e., análises internas. As perspectivas externas estão

atreladas às expectativas, e direito à informação, que os impactados desfrutam, i.e., análises externas.

Outros stakeholders desejarão informações de outra espécie, muitas das quais podem estar relacionadas a custos externos. Um bom exemplo aqui envolve as expectativas de esclarecimentos sobre os algoritmos de precificação de corrida empregado por empresas de transporte de passageiros, e.g., Uber, Cabify, 99, Easy, táxi, Wappa, Lyft etc. Os investidores querem saber quais empresas lucram mais por corrida, as empresas intencionam calcular o preço ideal que as permita o máximo de lucro dado o valor cobrado pela concorrência, os passageiros vão indagar qual o menor valor possível pela corrida, os motoristas desejam ponderar qual bandeira lhes remunera melhor e gestores públicos podem manifestar interesse por evitar práticas nocivas, e.g., tarifas abusivas em horário de pico.

Historicamente, a maioria das externalidades criadas foi ignorada (KOLSTAD, ULEN e JOHNSO, 2003). Porém, os gestores precisam coletar os dados pertinentes para capacitá-los a gerenciar os impactos da IA. O relato nem sempre é direto, dado que determinar a extensão e a respectiva atribuição dos custos costuma ser complicado, mas a divulgação adequada ajuda a tornar a propalação da prestação de conta como uma prática profícua e meritória.

Salientemos que algumas externalidades podem ser vistas como positivas (benefícios) e elogiadas. Quando consideradas significativas, devem ser relatadas, já que as partes afetadas têm o direito de saber sobre os impactos potencialmente significativos. Citamos dois exemplos da Tesla. No primeiro, a empresa melhora o desempenho das portas do Modelo X com um update via software (CARDOSO, 2016). O segundo, consiste numa atualização, denominada “Modo cão”, que ajusta a temperatura do interior do veículo desligado para que o ambiente fique seguro para os animais que estão à espera de seus donos, além de exibir a mensagem “Meu dono volta em breve” e a temperatura no interior do veículo (RODRIGUES e MU, 2019).

No intuito de estabelecer quais tipos de informações serão coletadas, nos valemos das respostas obtidas nas questões QN1 e QN2 (i.e., por que co-

letar e divulgar e quem contemplar, respectivamente) para estabelecermos a finalidade e o objeto da nossa ação. Da terceira, QN3 (i.e., o que coletar e divulgar), dimana a perquirição de onde obter o que será mensurado.

É bem verdade que já vislumbramos nos itens 6.2 Onde coletar e 6.3 Como coletar, de que modo proceder para apurar informações ex-ante e ex-post sobre a IA. Contudo, pelo exposto acima, faz-se necessária uma ampliação do escopo de nossa sondagem. Destarte, averiguemos a influência da cadeia de suprimentos, no ofício de dimensionar.

7.2.2 Onde medir

Tão importante quanto saber o que divulgar é saber onde obter as informações que se quer propalar. No item 6.2 Onde coletar, estudamos todo o processo de construção da IA. Agora, precisamos ampliar nosso escopo e compreender as origens dos recursos utilizados pela IA (i.e., cadeia de suprimentos) e investigar, de forma ampliada, os resultados dessas provisões, quer para sua confecção quer para sua operação, através da análise do ciclo de vida.

7.2.2.1 Cadeia de suprimentos

Uma cadeia de suprimentos pode ser definida como a rede entre uma organização e seus fornecedores e clientes, conforme necessário para produzir e distribuir um bem ou serviço específico. As empresas frequentemente terceirizam aspectos de suas operações para outras organizações não relacionadas, essas passam assim a compor a cadeia de suprimentos como fornecedores (VITASEK, 2019).

No caso da inteligência artificial não é diferente, especialmente quando embarcada (e.g., ilustração 2 ou em um telefone). Na ilustração 2 (veículo inte-

ligente) a percepção de que várias partes são produzidas por diferentes atores e combinadas num produto final (i.e., como uma agremiação) remonta o final do século XIX. Mais especificamente, o produtor da IA que “percebe” o ambiente ao redor do carro, não é o mesmo que fabricou os sensores empregados na captação dessas informações e pode não ser o mesmo que produziu a IA que “dirige” o veículo (ou os dados em que foi “treinada”), entretanto, todos esses atores concorrem para os mesmos objetivos.

Não obstante, na ilustração 2, acidente que envolveu um Tesla Model 3 que bateu no teto de um caminhão tombado na pista (記者林宜樟, 2020), estabelecer o culpado pode não ser tão simples. O sistema da Tesla usa principalmente câmeras para se orientar e geralmente, objetos imóveis são um entrave. Quem causou o acidente: a IA que não “viu” ou a câmera que não “captou” o caminhão? Uma alternativa possível seria o uso de LIDAR (Light Detection And Ranging, i.e., detector fotométrico) que forneceria melhores informações de distância para objetos imóveis. É preciso ressaltar, contudo, que o motorista deve sempre estar atento ao trajeto e que poderia ter evitado o acidente.

Para o exemplo do telefone, mesmo que deixemos as questões de hardware de lado, a coexistência de aplicativos, nativos e de terceiros, eleva a percepção de agremiação a patamares de ecossistema. Se nos determos apenas aos dados, consumidos e gerados pelo aparelho e por terceiros, no caso de um vazamento de dados, quem podemos culpar: o fabricante do aparelho, do sistema operacional, do aplicativo, ou mesmo o Google ou o Facebook (cujas APIs coletam dados de forma transparente)?

Para dirimir essas e outras questões precisamos entender como essas matérias-primas são processadas (i.e., por quem, como, quando, onde etc.). Assim lançamos mão da técnica de Análise do Ciclo de Vida para auxiliar na coleta dos dados que embasarão nossas investigações.

7.2.2.2 Análise do Ciclo de Vida (ACV)

Se os intendentes desejam compreender, e potencialmente medir e controlar, os impactos econômicos, sociais e ambientais relevantes criados por seus produtos e serviços, a análise do ciclo de vida (ACV) é uma forma de obter esse conhecimento.

ACV é o processo de análise ou avaliação de um produto ou serviço ao longo de toda a sua vida, geralmente referido como análise de um produto ou serviço desde o berço até o túmulo (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2006). Por exemplo, para a IA, a ACV pode começar com a análise dos objetivos e requisitos do projeto de uma perspectiva de negócios, convertendo esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para atingir os objetivos e terminar com uma consideração do que acontece com o produto no final de sua vida; i.e., seu despojo. Diferentes colaboradores decidiram sobre a extensão e alcance da existência analisada, bem como os tipos e impactos aferidos.

A referida inspeção fornece aos gestores informações que lhes permite tomar decisões sustentáveis e responsáveis. Uma vez de posse desse conhecimento, eles podem considerar quais melhorias podem ser necessárias, quais etapas de produção podem ser alteradas e assim por diante. Os impactos classificados como importantes são priorizados para inspeção e melhoria.

Como já discutido, as partes envolvidas possuem interesses diversos que devem ser contemplados. Por isso, nem toda informação apresentada será de natureza numérica, e.g., visões, decisões, metas e políticas empregadas também devem integrar os relatos apresentados. Os resultados obtidos podem ser, e alguns serão, publicizados (e.g., por meio de informes externos) e a análise pode ser performada, ou os resultados obtidos confirmados, por terceiros, geralmente sobre a forma de auditoria externa.

Sob a perspectiva de *accountability* (especialmente no tocante a quais divulgações fazer) ao projetar uma imagem pública de aceitação da responsabilidade pelos impactos ao longo do ciclo de vida de um produto, a organização

se diferencia de seus concorrentes, que podem não demonstrar um nível tão alto de responsabilidade. Além disso, quanto mais informados os diferentes stakeholders (e.g., consumidores, investidores, credores, funcionários, governo, mídia) se tornam sobre os ciclos de vida dos diferentes produtos e serviços, maior será sua capacidade de fazer uma avaliação informada e escolherem sobre quais organizações suportar.

Salientemos aqui que quase sempre o período estudado será pontual ou alcançará algumas etapas do ciclo de vida. Contudo, sem o conhecimento da localização do episódio na totalidade, não seremos capazes de deslindar o liame que interliga causa e efeito ou responsável e encargo.

Uma vez que esses fornecedores estão provendo insumos para atender às demandas da IA, eles têm alguma responsabilidade pelos impactos mencionados acima? Apenas por meio da cadeia de suprimentos e da Análise do Ciclo de Vida ajuntamos relações e não causalidades.

Devemos levar nossa análise ainda mais longe e obter provas (i.e., dados e outros artefatos) para embasar a relação de causa e efeito. Para tal, apoiamo-nos no conceito contábil ampliado de custo e na Análise de Custo do Ciclo de Vida.

7.2.3 Como medir

Carecemos de um método que analise os valores dos recursos consumidos, ou impactados, como resultado da operação de um curso de ação e seja expresso em termos quantitativos ou qualitativos (i.e., atribui uma valoração). Também deve incluir a determinação dos ganhos e perdas que poderiam ocorrer ao se tomar outro curso de ação (i.e., comparação) ou selecionar a alternativa mais viável (i.e., otimizar). Além disso, ser aplicável a uma instância, parte ou a todo o processo analisado. Para tal, aplicaremos o conceito ampliado de custo.

7.2.3.1 Custo

Medida, expressa em termos financeiros ou não financeiros, do valor dos recursos consumidos ou impactados como resultado das operações de uma organização. Pode incluir tanto custos privados quanto custos externos (O'SULLIVAN e SHEFFRIN, 2003). Vamos nos valer da mesma definição para expressar os custos da IA.

Custos privados são os custos que o comprador de um bem ou serviço paga ao vendedor. Os custos externos (muitas vezes referidos como externalidades) podem ser definidos como: os impactos que uma entidade tem sobre as partes externas à organização, onde esses grupos não são os compradores ou vendedores de bens ou serviços específicos nem concordaram ou participaram nas atividades que causaram a externalidade (KOLSTAD, ULEN e JOHNSO, 2003).

Por exemplo, o custo de fabricação de uma IA (e.g., os custos com servidor, armazenamento, segurança e mão de obra) reflete o custo privado para o fabricante. Vazamento de dados ou erros de classificação, retratam externalidades imputadas diretamente aos afetados pelas exposições ou equívocos. Contudo, podem estimular males econômicos, sociais e ambientais de monta coletiva, e.g., a contenda envolvendo avaliadores de risco como o COMPAS.

Embora os custos externos possam ser difíceis de quantificar, ainda podemos fornecer uma descrição dos impactos que as operações da IA teve em um evento específico, mesmo que não possamos atribuir um valor a tais custos. Mas, o que reconhecer como custo?

Como discutido anteriormente, isso vai depender de quais obrigações e esclarecimentos temos antelação. De mais a mais, as diferentes partes interessadas mantêm desvelo ou necessidades informacionais distintas. Verbi gratia, citando apenas duas classes, no caso dos preditores de reincidência criminal, a sociedade como um todo possui interesse nas taxas de acerto e erro, uma vez que não quer encarcerar um inocente nem alforriar um culpado, e os pacientes

possuem interesse nas métricas e critérios de imparcialidade empregados na análise, a fim de alegar iniquidade.

Um ponto importante a ser destacado aqui é que, havendo mais de uma maneira de mensurar o uso de qualquer recurso, é importante que se identifique qual estrutura assinalada para fazer suas mensurações, i.e., metodologia e métodos empregados. Estruturas diferentes podem, potencialmente, gerar medições diferentes, e.g., no caso do algoritmo de corte de imagem do Twitter que fora testado para imparcialidade, mas falhou na contestação pública.

Em não havendo estrutura comum para o custo em exame, pode-se optar por fornecer uma descrição clara dos seguintes aspectos:

• O que causou [...] e o que foi feito para impedir [...];

• Quais são os custos [...] associados à [...];

• Quais danos ocorreram a curto e longo prazo (sendo muito claro sobre os vários tipos de impactos negativos);

• As ações que estão sendo tomadas para reduzir esses impactos negativos;

• As ações tomadas para garantir que este tipo de evento não aconteça novamente;

• O que aprendeu com o incidente;

• Quais procedimentos operacionais falharam;

• Como seus procedimentos de resposta emergencial funcionaram ou não.

Idealmente o melhor momento para se reportar uma falta é antes que ela aconteça, i.e., por meio de relatórios e ações de prevenção empregadas na obstrução da anomalia. Como nem todos os imprevistos são previsíveis ou relevantes, consagra-se reportar o incidente assim que ocorra, e encerra seu acompanhamento apenas depois de sanado. Detalhes de seus procedimentos de resposta emergencial também devem estar disponíveis, de forma que as

partes interessadas possam avaliar os riscos decorrentes das atividades em andamento.

A seguir, aplicaremos a noção ampliada de custo ao ciclo de vida. Dessarte, seremos capazes de aferir impactos criados pelas operações da organização (e da IA), atribuir tais custos a atividades e produtos específicos e selecionar alternativas mais viáveis entre as possíveis.

7.2.3.2 Análise de Custo do Ciclo de Vida (ACCV)

A Análise de custo do ciclo de vida (ACCV) procura colocar um custo nas várias entradas (recursos usados), saídas (incluindo resíduos e emissões) e impactos criados pelas operações da organização, e atribuir tais custos a atividades e produtos específicos (KLOEPFFER, 2008). É uma ferramenta de gestão que auxilia a organização a selecionar a alternativa mais viável entre as possíveis, uma vez que são considerados os custos de compra, operação, manutenção e, em última instância, descarte de um produto.

Análise parecida pode ser empregada no ciclo de vida da IA para computar os diversos custos associados, face uma ênfase específica (e.g., imparcialidade ou explicabilidade) atribuindo tais custos a atividades ou produtos específicos. Empiricamente, essa prática é executada por diversos desenvolvedores ao treinar uma miríade de modelos em paralelo e observar aqueles que apresentam melhores resultados, e.g., preditivos (i.e., ênfase nos índices de acerto, erro etc.), menores consumo energético (e.g., muito importante em sistemas embarcados) ou melhores avaliações de imparcialidade (i.e., ênfase nos índices de discriminação social, etário, racial etc.), bem como cada etapa do processo contribuiu, e custou, e quais, eventualmente, podem ser otimizadas.

À primeira vista, lobrigam-se as semelhanças entre as referidas técnicas e métodos analíticos como *Failure Mode and Effects Analysis* (FMEA) e *Failure Mode Effects and Criticality Analysis* (FMECA). A FMEA fornece uma avaliação dos modos de falha em potencial para projetos e processos e seu provável

efeito nos resultados e desempenho do produto. Uma vez que as circunstâncias de falha são estabelecidas, o controle de risco pode ser aplicado para eliminar, conter, reduzir ou controlar as falhas potenciais. Ao se incorporar ao estudo, uma investigação da gravidade das consequências, suas respectivas probabilidades de ocorrência e sua detectabilidade, obtemos o FMECA. Para que tal análise seja realizada, as especificações do produto ou processo devem ser estabelecidas e seu emprego pode identificar locais onde ações preventivas adicionais podem ser apropriadas para minimizar riscos.

Símile para os demais artifícios praticados em disciplinas como segurança, qualidade, imparcialidade etc. Todas essas ferramentas possuem o mesmo objeto de estudo, i.e., o ciclo de vida do produto ou serviço, porém, sobre concepções diferentes, e.g., custo, segurança, qualidade etc. Isso quer dizer que quando respondemos as questões normativas QN3 e QN4 (i.e., sobre coleta e divulgação de informação), estamos diretamente, ou indiretamente, escolhendo quais instrumentos nos valer.

Contudo, pode ser muito confuso, ou mesmo impraticável, implementar uma dessas metodologias, se, no início da análise, houver muitos aspectos sendo tratados. Pode ser difícil atribuir um valor (e.g., custo ou probabilidade) às externalidades, mas elas devem ser identificadas e, quando possível, descritas tanto em termos qualitativos quanto por meio das várias formas de mensurações quantitativas cabíveis. As informações sobre externalidades fazem parte do cômputo geral de informações sobre as quais as decisões informadas subsequentes são tomadas.

Até aqui, vimos as principais técnicas gerais de onde e como medir os diversos custos privados e externalidades. No tópico seguinte, 7.3 XAI, nos aprofundaremos nas técnicas de Explainable Artificial Intelligence. Esses dispositivos são análogos aos estudados anteriormente tanto pelo objeto de análise (i.e., recaem sobre a cadeia de suprimentos e o ciclo de vida e analisam custos específicos) quanto pela finalidade (e.g., auxiliar a organização a selecionar a alternativa mais viável entre as possíveis, identificar locais onde ações

preventivas adicionais podem ser apropriadas para minimizar riscos etc.) só que próprios para escrutinar a IA.

7.3 XAI

Em 2019, Gunning declarou:

> *"A XAI criará um conjunto de técnicas de aprendizado de máquina que permite que usuários humanos entendam, confiem adequadamente e gerenciem com eficácia a geração emergente de parceiros artificialmente inteligentes" (Gunning, 2019).*

A inteligência artificial explicável (IAE ou XAI) é um conjunto de processos e métodos que permite os usuários humanos compreenderem e confiarem nos resultados e saídas criados por algoritmos de aprendizado de máquina (IBM, 2021).

A IAE é usada para descrever um modelo de IA, seu impacto esperado e possíveis vieses. Além de ajudar a caracterizar a precisão, imparcialidade, transparência e explicar os resultados da IA.

Talvez a forma mais natural de explicação seja o exemplo. Nesse contexto, XAI se concentra em extrair espécimes representativos que compreendem as relações internas, peculiaridades e correlações externas de um modelo ou previsão para entender melhor por que isso aconteceu, (i.e., porque o modelo se comporta dessa maneira em um contexto específico).

Então, que tipos de perguntas o XAI responde? Em sua grande maioria, ela tenta responder a perguntas "refutáveis". Uma hipótese é falsificável (ou refutável) se puder ser logicamente contrariada por um teste empírico (e.g., que pode ser executado com as tecnologias existentes) (ANGELES, 1992). Por

exemplo, "[uma certa condição] levou a IA a se comportar mal?". Esse exemplo deve ter uma resposta testável (i.e., sim ou não).

Outro exemplo muito corriqueiro são as perguntas políticas, (i.e., centrados em valores que não são falsificáveis). Por exemplo, “Deveria haver uma lei sobre isso” ou “Deveríamos ter uma IA imparcial”. Embora possam ser endereçáveis, XAI visa nos ajudar a ver como as coisas são e não como deveriam ser, mesmo que às vezes lide com a última.

Por outro lado, uma questão normativa é do tipo: "O que deveria ter acontecido?" ou "Teria uma pessoa se comportado dessa maneira?". A resposta para essas duas depende de comparação (i.e., contexto): a primeira com um cenário plausível, a segunda com um comportamento humano. Mais especificamente, a primeira é o objetivo de uma técnica XAI chamada contrafactual enquanto a segunda é o objeto de estudo numa área específica da IA, a saber, IA Moral.

Para esclarecer ou exemplificar um determinado comportamento ou funcionalidade, investigar as consequências da IA que causam danos à indivíduos ou enfrentar as questões relacionadas à responsabilidade legal, é importante fornecer clareza sobre como uma decisão algorítmica foi tomada. Assim, atenção especial é dada ao tópico da explicabilidade e ao relacionado (embora distinto) tópico da inteligibilidade.

7.3.1 Explicabilidade

A IA precisa ser explicável, porque a explicabilidade é uma ferramenta crítica na construção de confiança pública e na compreensão da tecnologia. Por explicabilidade, queremos dizer uma combinação do sentido epistemológico de “inteligibilidade” (i.e., como funciona) e o sentido de “*accountability*” (i.e., quem é o responsável). Uma explicação adequada deve abranger ambos.

Por meio de técnicas e métodos XAI, podemos fornecer os meios necessários para explicar elementos do “raciocínio” que levaram uma máquina a

tomar uma determinada decisão e o processo nela contido (Gunning & Aha, 2019). No contexto da IA, a interpretabilidade pode ser definida como "a capacidade de explicar ou esclarecer para um humano em termos compreensíveis" (DOSHI-VELEZ e KIM, 2017).

A necessidade de interpretabilidade surge de uma incompletude na formalização do problema (i.e., para certos problemas ou tarefas não é suficiente obter apenas a previsão). Por isso, o modelo também deve explicar como chegou à previsão, dado que um prognóstico correto resolve parcialmente o seu problema original (DOSHI-VELEZ e KIM, 2017).

Assim, um sistema de IA dotado de interpretabilidade, é aquele capaz de averbar alguma aclaração sobre seus cômputos (i.e., como os cômputos foram gerados). Observe que a interpretabilidade é sobre ser capaz de discernir a mecânica sem necessariamente conhecer a causa (i.e., como funciona), e a explicabilidade refere-se ao desvelar das razões (i.e., porque funciona).

7.3.1.1 Taxonomia dos Métodos de Explicabilidade

As técnicas de XAI nos permitem diferentes graus de análise. Podemos classificar os modelos de acordo com sua opacidade (ou seja, falta de interpretabilidade). Os modelos de caixa branca (interpretáveis) podem ser claramente explicados em termos de como se comportam, como produzem previsões e as variáveis de influência. Esses modelos incluem modelos baseados em regras, árvores de decisão e regressões lineares.

Com modelos caixa-preta, os usuários podem explicar a relação entrada-saída, mas as razões subjacentes ou processos envolvidos na produção da saída não estão disponíveis (i.e., explicitamente, como em um modelo caixa branca). Os modelos de caixa preta geralmente resultam em maior precisão quando comparados aos modelos de caixa branca, mas sacrificam a explicabilidade. As caixas pretas estão no centro de nosso estudo daqui para frente.

A Tabela 9 – Árvore de explicações – ilustra as diversas fontes de nossas explicações. Cada folha da árvore representa a fonte e o tipo de questão normativa que explica. Embora vitais, as técnicas de visualização foram omitidas porque se encaixam em todas as categorias.

Conforme ilustrado na Tabela 9 – Árvore de explicações, as fontes primárias de nossos aclaramentos serão os dados e o modelo. Essas matrizes poderão ser usitadas em dois momentos distintos: ex-ante e ex-post. Ao analisarmos os dados de treinamento (ex-ante), podemos entender as características e atributos (i.e., as variáveis correlacionadas) desses dados antes de iniciar o treinamento, identificar variáveis irrelevantes, descobrir ou verificar relacionamentos importantes que os modelos de aprendizado de máquina devem refletir, e remover os vieses subjacentes aos dados antes de qualquer modelagem. Numa análise ex-post, geralmente via dados da XAU, extraímos um conjunto cronológico de entradas e saídas (i.e., a cadeia de eventos que levam a um ato) fornecendo uma interpretação bruta do que aconteceu e um conjunto factual para indagações posteriores ao modelo. As técnicas empregadas na explicação dos dados são abordadas no item 7.3.1.2.1 Dados.

Tabela 9 – Árvore de explicações

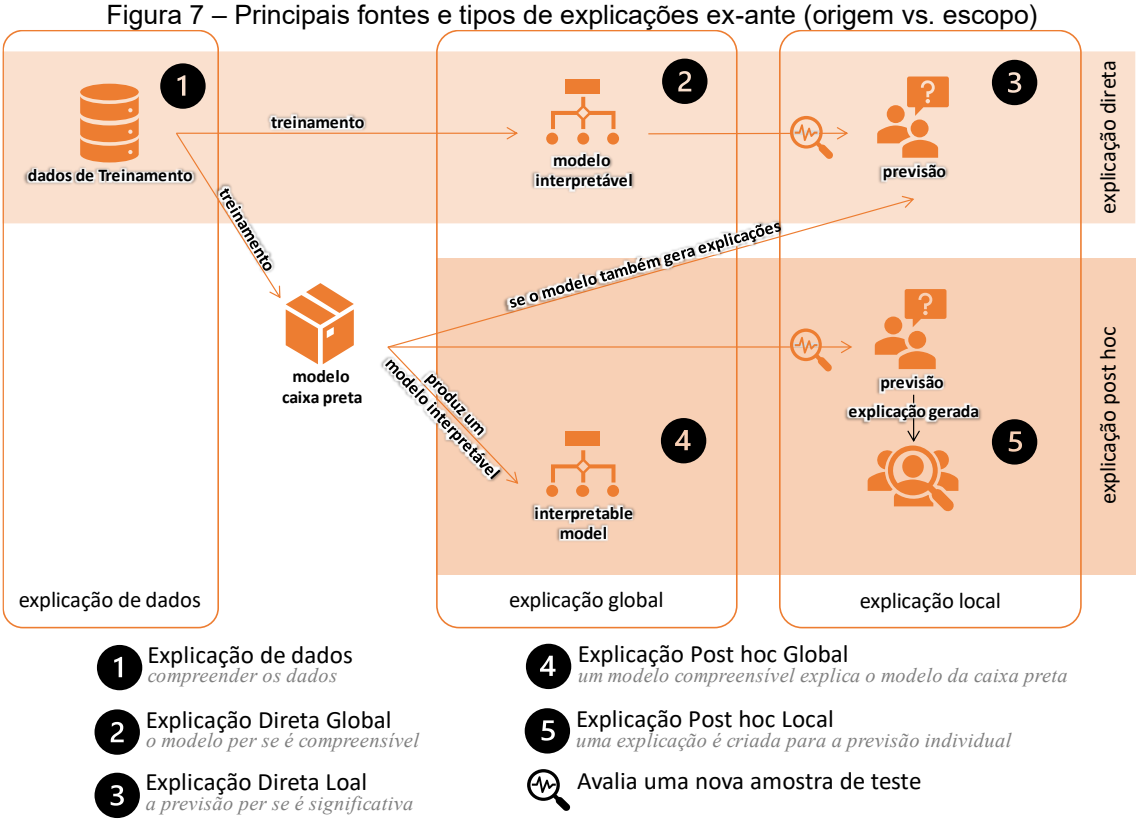
Podemos extrair nossas explicações do	modelo	via explicação...	global	usando ...	Explicação Post-hoc: Construída sobre modelos de caixa preta para fornecer compreensão global Modelo Interpretável: Oferece segurança, confiabilidade e conformidade
			local	baseada em...	Explicação extraída: Explicações extraídas usadas para prever novas amostras Características: Destaca características que estão necessariamente presentes ou ausentes para que a previsão ocorra Exemplos: Raciocínio baseado em casos usando protótipos e críticas[55]
	dados			Baseadas em...	Características: Representações desembaraçadas fornecem recursos significativos Exemplos: Raciocínio baseado em casos usando protótipos e críticas

Fonte: Mojsilovic (2021)

[55] Vide A.2.3.2 Protótipos e Críticas.

Em relação ao modelo, estaremos sempre preocupados em entender como ele funciona sob dois pontos de vista, global (holístico) e local (pontualmente), por meio de explicações, características e amostras. Analisando o modelo em treinamento (ex-ante), podemos investigar, por meio das técnicas de XAI, questões sobre qualidade, segurança, proteção, robustez, transparência e assim por diante. Os resultados dessas análises serão cruciais para sustentar um caso de previsibilidade, por exemplo. A utilização de técnicas de XAI como um conjunto de ferramentas forenses (uso ex-post) permite-nos explicar as suas deliberações e comportamentos, em geral ou em situações específicas, durante a investigação. As técnicas empregadas na explicação do modelo são abordadas em 7.3.1.2.2 Modelo.

Figura 7 – Principais fontes e tipos de explicações ex-ante (origem vs. escopo)



Fonte: Mojsilovic (2021)

A Figura 7 – Principais fontes e tipos de explicações ex-ante ilustra as principais fontes de explicação ex-ante (i.e., durante as fases 2 a 5, Compreensão dos dados, Preparação dos dados, Modelagem e Avaliação, segundo o

CRISP-DM). Nela, podemos ver que o processo de explicação começa pelos dados. Uma vez treinado o modelo, obtemos explicações sobre seu funcionamento e os classificamos de acordo com sua origem (direta e post hoc) e escopo (global e local).

A seguir, examinamos essas categorias em detalhes. Ilustramos cada uma delas com algumas de suas principais técnicas, sempre abordando os seguintes critérios: (1) origem, (2) resultado, (3) especificidade e (4) escopo.

(1) Quanto a origem: em dados, direta (intrínseco ao modelo) ou post hoc (após aplicação do método);

(2) Quanto ao resultado: relatório estatístico dos atributos, visualização do relatório de atributos, mecanismo do modelo, pontos de dados (vetores) e modelo intrinsecamente interpretável;

(3) Quanto a especificidade: em agnóstico ou genérico (aplicável a qualquer modelo) ou específico (limitado a um modelo ou classe);

(4) Quanto ao escopo: em local ou global.

7.3.1.1.1 Quanto a origem

Este critério trata da fonte da interpretabilidade. Por origem, temos os dados usados para treinar o modelo, o próprio modelo como fonte (denominada de direta, intrínseca, específica, própria ou *intrinsĕcus*) ou aplicando métodos que analisam o modelo após o treinamento (denominada indireta, pós, posterior ou *post hoc*).

Direta – refere-se a modelos de aprendizado de máquina que são considerados interpretáveis devido à sua estrutura simples, e.g., árvores de decisão curtas ou modelos baseados em regras.

Indireta – refere-se à aplicação de métodos de interpretação após o treinamento do modelo, e.g., valores Shapley ou modelos substitutos.

7.3.1.1.2 Quanto ao resultado

Os métodos de interpretação indireta podem ser classificados de acordo com a natureza dos resultados produzidos.

Relatório Estatístico dos Atributos – vários métodos de interpretação apresentam estatísticas para cada atributo. Alguns métodos retornam um simples número por atributo ou um resultado mais complexo.

Visualização do Relatório de Atributos – a maioria das informações contidas no relatório estatístico dos atributos pode ser visualizada. De fato, alguns atributos são realmente significativos apenas quando visualizados. Por exemplo, t-SNE e grafos de correlação.

Mecanismo do Modelo – aqui nos interessa como o modelo funciona, i.e., seu maquinismo. Por exemplo, numa árvore de decisão cada nó interno testa um atributo, cada ramo corresponde a um valor do atributo e cada folha atribui uma classificação. Abarca essencialmente os modelos de interpretabilidade direta, mas não se restringe apenas a estes. Estes métodos de interpretabilidade que geram o mecanismo do modelo são tidos como específicos.

Pontos de Dados (vetores) – esta categoria inclui todos os métodos que retornam pontos de dados (existentes ou avaliados) para tornar o modelo interpretável. Um ponto de dados, vetor ou observação é um conjunto de uma ou mais medições sobre uma única previsão. Para serem úteis, esses métodos exigem que os próprios pontos de dados sejam interpretáveis. Isso funciona bem para imagens e textos, mas são menos adequados para pontos com centenas de atributos. Exemplos incluem: identificação de protótipos de classes previstas[56] e explicações contrafactuais, i.e., o método encontra um ponto de dados semelhante,

[56] Essas arquiteturas contêm um auto codificador e uma camada especial de protótipo. Cada unidade dessa camada armazena um vetor de peso que se assemelha a uma entrada de treinamento codificada. Como os protótipos aprendem durante o treinamento, a rede neural resultante contém as explicações para cada previsão sua, e essas explicações são fiéis ao que a rede realmente computa.

alterando alguns atributos para os quais o resultado previsto é alterado significativamente.

Modelo intrinsecamente interpretável (modelo substituto) – consiste em aproximar (local ou globalmente) os modelos caixa-preta a modelos interpretáveis. O modelo interpretável é entendido observando-se o mecanismo do modelo ou os relatórios estatísticos dos atributos.

7.3.1.1.3 Quanto à especificidade

A maioria das técnicas de XAI que cobrimos pode ser aplicada a qualquer classe de modelo, daqui em diante modelo agnóstico ou genérico. No entanto, algumas dessas técnicas têm restrições de aplicabilidade e só podem ser aplicadas a uma classe de modelo específica para obter uma explicação, doravante denominada modelo específico.

Genérico ou Agnóstico – podem ser aplicados a diferentes tipos de modelos. Por exemplo, a técnica LIME (vide A.2.2.3 Modelo Substituto Local) pode ser usada para interpretar praticamente qualquer conjunto de entradas e previsões de aprendizagem de máquina. Embora essas técnicas sejam convenientes (i.e., devido a sua generalidade) elas geralmente se baseiam em modelos substitutos ou outras aproximações o que pode degradar a precisão das explicações fornecidas.

Específico – as ferramentas de interpretação específicas são limitadas a classes específicas ou a um único tipo de modelo (i.e., possuem uso restrito a certas tecnologias). Por exemplo, a técnica conhecida como *treeinterpreter* (vide A.2.2.5 Treeinterpreter) pode ser aplicada apenas a modelos de árvore de decisão. Essas técnicas de interpretação tendem a usar o modelo para ser interpretado diretamente, levando a explicações potencialmente mais precisas.

7.3.1.1.4 Quanto ao escopo (escala)

É importante compreender o funcionamento do modelo que você treinou como um todo (i.e., explicações globais) e ser capaz de entender como certos resultados foram obtidos (i.e., explicações locais). Ao examinar as relações entre as variáveis de entrada e saída de maneira funcional, os comportamentos locais ou globais tornam-se aparentes. Por exemplo, ao simular o seguro do seu carro, se mudarmos a motorização, o valor do seguro muda proporcionalmente (efeito global), enquanto mudar a cor de um carro produz pouca ou nenhuma mudança no valor do seguro, a menos que seja uma pintura premium (efeito local). Dessa forma, os métodos de interpretação podem ser classificados de acordo com o escopo (i.e., a abrangência da explicação) em local ou global.

Local – interpretações locais nos ajudam a entender as previsões do modelo para um único ponto de dados ou um grupo semelhante. Devido à proximidade desses pontos de dados, as explicações locais podem ser mais precisas do que as explicações globais.

Global – interpretações globais nos ajudam a entender as relações funcionais do modelo (i.e., explicações globais sobre o modelo) seus resultados ou o relacionamento entre as previsões e as variáveis de entrada. No entanto, as interpretações globais podem ser altamente aproximadas em alguns casos.

As melhores explicações sobre um modelo emanam da combinação dos resultados de técnicas de interpretação global e local.

Examinaremos agora as principais técnicas de IAE. São dissertados dois grupos: explicabilidade (i.e., técnicas que visam aclarar a mecânica da IA) e imparcialidade (i.e., técnicas que ambicionam dirimir vieses e injustiças nos algoritmos). Vale salientar que outros grupos de técnicas (e.g., robustez, moral, transparência etc.) existem e são imprescindíveis para dirimir questões sobre segurança, ética, recursos naturais etc.

7.3.1.2 Técnicas de Explicabilidade

As técnicas de explicabilidade serão classificadas quanto a origem, i.e., dados ou modelo (direta ou post hoc). Para cada classe serão apresentadas uma breve descrição e as principais técnicas. Resultado, especificidade, escopo bem como uma expectativa do nível de compreensão que cada técnica provê, serão abordados dentro de cada método no Apêndice 1. Uma categoria especial (dentro de modelo) intitulada “exemplos”, foi criada para compreender métodos que – apesar de poderem ser incluídos nas outras categorias – deveriam figurar juntos, devido a sua finalidade.

7.3.1.2.1 Dados

Os modelos de aprendizagem de máquina têm sua gênese nos dados usados em seu treinamento. Assim, é benéfico compreender as características e atributos desses dados antes de começar o treinamento. Para tal, veremos técnicas de visualização, compreensão e higienização desses dados.

Conjuntos de dados que contêm imagens, texto ou mesmo dados comerciais com muitas variáveis (i.e., alta dimensão) podem ser difíceis de visualizar como um todo (HINTON e SALAKHUTDINOV, 2006). Gráficos de dispersão são usados para apresentar os principais elementos estruturais de um conjunto de dados, e.g., clusters, hierarquia, escassez e outliers.

As projeções 2D permitem que esses conjuntos de dados de alta dimensão sejam projetados em espaços de baixa dimensão, representativos e visualizados usando a técnica de plotagem de dispersão (OSOWSKY, GAMBA, *et al.*, 2004). São frequentemente empregadas na detecção de fraudes ou anomalias para encontrar entidades periféricas, como pessoas, transações ou computadores, ou aglomerados incomuns de entidades (CASALS, 2017).

As principais técnicas de projeção 2D são: Non-Linear Principal Components Analysis (n-LPCA), Multidimensional Scaling (MDS), t-Distributed Sto-

chastic Neighbor Embedding (t-SNE), e Autoencoder Networks. Essas técnicas são apresentadas de forma resumida no item A.1 Explicação de dados do anexo 1.

7.3.1.2.2 Modelo

Como vimos em 7.3.1.1 Taxonomia dos Métodos de Explicabilidade, a fonte da explicação pode ser o próprio modelo (direta) ou resultado da aplicação de métodos que analisam o modelo após o treinamento (indireta). Uma terceira categoria, exemplos, foi criada para abarcar técnicas que abordam a explicabilidade de uma maneira mais compreensível. Os métodos de explicação baseados em exemplos selecionam instâncias específicas dos dados para explicar o comportamento dos modelos ou para explicar a distribuição de dados subjacentes.

7.3.1.2.2.1 Direta

Um critério importante para o sucesso de um projeto de IA é determinar o grau de interpretabilidade necessária. Se prover explicação é vital para o seu projeto, é melhor usar uma técnica de modelagem interpretável (i.e., devido à sua estrutura simples) desde o início.

As principais técnicas de explicação direta são: árvore de decisão, modelos baseados em regras e regressão linear. Essas técnicas são apresentadas de forma resumida no item A.2.1 Explicação direta do anexo 1.

7.3.1.2.2.2 Post hoc ou Indireta

A interpretabilidade indireta refere-se à aplicação de métodos de interpretabilidade ao modelo treinado para extrair explicações sobre seu funciona-

mento. Contemplaremos seis técnicas frequentemente citadas na literatura sobre o assunto: Individual Conditional Expectation (ICE), Modelo Substituto Global, Modelo Substituto Local, Regras de Escopo (Anchors), Treeinterpreter e Valores Shapley. Essas técnicas são apresentadas de forma resumida no item A.2.2 Explicação Post hoc ou indireta do anexo 1.

7.3.1.2.2.3 Exemplos

Os métodos de explicação baseados em exemplos selecionam instâncias específicas dos dados para explicitar o comportamento dos modelos ou para expor a distribuição de dados subjacente. Contudo, isso só faz sentido se pudermos representar uma instância dos dados de maneira compreensível.

Exemplos ajudam-nos a construir modelos mentais da IA e dos dados nos quais ela foi treinada. As ferramentas de explicações baseadas em exemplos, em geral, não são atreladas a um modelo específico, isso porque tornam qualquer IA mais interpretável.

A diferença para os métodos que vimos até aqui é que as ferramentas baseadas em exemplos explicam um modelo selecionando instâncias do conjunto de dados (i.e., um caso específico ou conjunto de) e não criando resumos de atributos ou listas. Cobriremos (de forma resumida no item A.2.3 Exemplos do anexo 1) os seguintes métodos de interpretação baseados em exemplo: contrafactuais, protótipos e críticas (MMD-critic e ProtoDash), instâncias influentes, diagnóstico por exclusão e funções de influência.

7.3.2 Imparcialidade

Imparcialidade é um aspecto importante da interpretabilidade e um objetivo para qualquer projeto de IA cujo resultado afetará vidas humanas. Os métodos para apurar justiça geralmente analisam as previsões do modelo para

segmentos demográficos sensíveis (e.g., etnia, gênero, cor, sexualidade), mas podem abordar casos individuais, i.e., verificar a inconformidade de uma previsão específica.

O estudo da neutralidade na IA progride rapidamente, incluindo o desenvolvimento de técnicas para: mitigar injustiças ou preconceitos nas previsões; e a criação e depuração de dados e modelos, a fim de torná-los mais justos.

Diferentes tipos de técnicas de justiça contemporâneas podem detectar tendências, corrigir vieses nas previsões do modelo e aprender a fazer previsões justas. Mas antes, precisamos saber quais os principais problemas que enfrentamos.

7.3.2.1 Problemas

Voltamos nossa atenção agora para as duas possíveis origens dos vieses na IA: dados e algoritmos. Segundo (**Unsupported source type (DocumentFromInternetSite) for source Exe16.**) as principais preocupações quanto a dados e algoritmos são:

> **Dados** – Mal selecionados, incompletos, incorretos ou desatualizados, selecionados com viés, perpetuam e promovem vieses históricos.
>
> **Algoritmos** – Sistemas mal projetados; serviços de personalização e recomendação que restringem em vez de expandir as opções do usuário; sistemas de tomada de decisão que assumem a ocorrência de correlações como existência de causalidades; algoritmos que não se resguardam contra vícios em seus dados de treinamento; e modelos não explicáveis ou não escrutinados e mitigados contra vieses.

Uma observação imediata é que uma IA é projetada para captar padrões estatísticos nos dados de treinamento. Se os dados de treinamento refletem vieses sociais existentes contra uma minoria, é provável que o algoritmo incor-

pore esses vieses. Isso pode levar a decisões desfavoráveis para os membros desses grupos.

A diferença na precisão da classificação entre os diferentes grupos é uma fonte importante e subestimada de injustiça. Assumindo um espaço de atributos fixo, um classificador geralmente melhora com o número de pontos de dados usados para treiná-lo.

Infelizmente, sempre há proporcionalmente menos dados disponíveis sobre minorias. Isso significa que nossos modelos, para minorias, geralmente tendem a ser piores que para a população em geral. A menos que ambos os grupos formem uma população homogênea, amostras adicionais podem beneficiar os dois grupos.

Diferenças culturais podem agravar os efeitos negativos das disparidades no tamanho da amostra. Isso nos leva a concluir que os padrões estatísticos que se aplicam à maioria podem ser inadequados para um grupo minoritário.

As previsões extraídas dos dados nem sempre são observadas na realidade. Entre as possíveis causas figuram: as variáveis podem não ter uma distribuição independente e gaussiana; as amostras podem ser tendenciosas; os rótulos nos dados podem estar incorretos; e os erros podem estar concentrados em uma classe específica.

Conseguir imparcialidade pode ser computacionalmente caro, principalmente quando colocamos demandas adicionais. Como algumas das aplicações mais interessantes da IA tendem a estar no limite do que é realmente computacional e humanamente viável, os recursos adicionais necessários para alcançar a justiça podem ser limitados.

Lograr neutralidade no âmbito jurídico poder ser tão improbo quanto na esfera técnica. Legislações antidiscriminatórias normalmente buscam proporcionar acesso igual a emprego, condições de trabalho, educação, proteção social, bens e serviços. Essas leis são muito diversas e incluem muitos conceitos legais como: requisitos profissionais legítimos; impacto separado e tratamento

díspar; ónus da prova e teste situacional; e o princípio da sub-representação de grupo.

A legislação atual oferece vários níveis de proteção contra: discriminação por pertencer a uma classe específica de gênero, idade, etnia, nacionalidade, deficiência, crenças religiosas e orientação sexual; tratamento desigual (i.e., tratamento depende do concernimento à classe); e impacto desigual (i.e., o resultado depende do concernimento à classe mesmo que, aparentemente, as pessoas sejam tratadas da mesma maneira).

Mas provar que a discriminação ocorreu é difícil. Em casos muito raros, pode haver uma confissão de que alguém discriminou outra pessoa porque era mulher, negra, homossexual etc. Em casos raros, as evidências são quase comicamente esmagadoras[57], mas na maioria dos casos, é difícil provar intenções. O ônus da prova pode ser compartilhado (como no Tribunal de Justiça Europeu) em casos de discriminação, i.e., o acusador deve apresentar evidências das consequências, o réu deve apresentar evidências de que o processo foi justo.

7.3.2.2 Métricas

(VERMA e RUBIN, 2018) coletaram as definições de imparcialidade mais importantes para o problema de classificação algorítmica, explicando a lógica por trás dessas definições. A análise também evidenciou os motivos pelo qual um mesmo caso pode ser considerado justo de acordo com algumas definições e injusto de acordo com outras.

São apresentados 5 grupos de métricas estatísticas: definições baseadas no resultado previsto (2); definições baseadas em resultados reais e previstos (7); definições baseadas em probabilidades previstas e resultados reais (4); medidas baseadas em similaridade (3); e raciocínio causal (4).

---

[57] Vide (King v Great Britain China Centre, 1991).

Definições baseadas no resultado previsto (2) - Focam nos resultados previstos para as várias distribuições demográficas e representam a noção mais simples e intuitiva de justiça.

Justiça de grupo (paridade estatística, taxa de aceitação igual, benchmarking). Um classificador satisfaz essa definição se indivíduos em grupos protegidos, e não protegidos, tiverem a mesma probabilidade de serem atribuídos à classe de previsões positivas (DWORK, HARDT, *et al.*, 2012).

Paridade estatística condicional. Essa definição estende a anterior, permitindo que um conjunto de atributos legítimos afete o resultado (CORBETT-DAVIES, PIERSON, *et al.*, 2017).

Definições baseadas em resultados reais e previstos (7) - Consideram as previsões para as diferentes distribuições demográficas e as compara com os resultados reais registrados no conjunto de dados.

Paridade preditiva (teste de resultado). Um classificador satisfaz essa definição se os grupos protegidos e não protegidos tiverem um valor preditivo positivo igual (CHOULDECHOVA, 2016).

Equilíbrio da taxa de erro falso positivo (igualdade preditiva). Um classificador satisfaz essa definição se os grupos protegidos e não protegidos tiverem índices de falso positivo iguais (CHOULDECHOVA, 2016).

Equilíbrio da taxa de erro falso negativo (oportunidade igual). Um classificador satisfaz essa definição se os grupos protegidos e não protegidos tiverem índices de falso negativo iguais (CHOULDECHOVA, 2016).

Chances equalizadas. Essa definição combina as duas anteriores: um classificador satisfaz a definição se os grupos protegidos e desprotegidos tiverem índices de verdadeiro positivo e índices de falso positivo respectivamente iguais (HARDT, PRICE e SREBRO, 2016).

Igualdade de precisão de uso condicional. Semelhante à definição anterior, essa definição conjuga duas condições: valores preditivos positivos e valores preditivos negativos iguais (BERK, HEIDARI, *et al.*, 2017).

Igualdade de precisão geral. Um classificador satisfaz essa definição se os grupos protegidos e não protegidos tiverem igual precisão de previsão (BERK, HEIDARI, *et al.*, 2017).

Igualdade de tratamento. Essa definição analisa a proporção de erros que o classificador faz e não a sua precisão. Um classificador satisfaz essa definição se os grupos protegidos e não protegidos tiverem uma proporção igual de falsos negativos e falsos positivos (BERK, HEIDARI, *et al.*, 2017).

Definições baseadas em probabilidades previstas e resultados reais (4) - Consideram o resultado real e o escore da probabilidade prevista.

Teste de Imparcialidade (calibração, correspondência de frequências condicionais). Um classificador satisfaz essa definição se, para qualquer escore de probabilidade prevista, os indivíduos nos grupos protegido e não protegido tiverem igual probabilidade de pertencer verdadeiramente à classe positiva (CHOULDECHOVA, 2016).

Boa Calibração. Essa definição estende a anterior, afirmando que, para qualquer escore de probabilidade prevista, os indivíduos nos grupos protegido e não protegido devem ter não apenas uma probabilidade igual de pertencer verdadeiramente à classe positiva, mas essa probabilidade deve ser igual à probabilidade prevista para uma determinada classificação (KLEINBERG, MULLAINATHAN e RAGHAVAN, 2017).

Equilíbrio da classe positiva. Um classificador satisfaz essa definição se os sujeitos que constituem a classe positiva dos grupos protegidos e desprotegidos tiverem o mesmo valor médio de probabilidade previsto (KLEINBERG, MULLAINATHAN e RAGHAVAN, 2017).

Equilíbrio da classe negativa. Em uma versão invertida da definição anterior, essa definição afirma que os sujeitos que constituem classe ne-

gativa de ambos os grupos protegidos e não protegidos também devem ter um score de probabilidade prevista, igual à média da probabilidade prevista para uma determinada classificação (KLEINBERG, MULLAINATHAN e RAGHAVAN, 2017).

Medidas baseadas em similaridade (3) – As definições estatísticas ignoram amplamente os demais atributos do sujeito classificado, exceto o atributo protegido ou sensível para o qual a não discriminação deve ser estabelecida. Essa abordagem pode ocultar diversas formas de injustiça.

Discriminação causal. Um classificador satisfaz essa definição se produzir a mesma classificação para quaisquer dois sujeitos com exatamente os mesmos atributos, i.e., todos os atributos adicionais que descrevem o indivíduo (GALHOTRA, BRUN e MELIOU, 2017).

Justiça através do desconhecimento. Um classificador satisfaz essa definição se nenhum atributo sensível for explicitamente usado no processo de tomada de decisão (KUSNER, LOFTUS, *et al.*, 2017).

Justiça através da conscientização. Essa definição é uma versão mais elaborada e genérica das duas anteriores: aqui, a justiça é capturada pelo princípio de que indivíduos semelhantes devem ter uma classificação semelhante (DWORK, HARDT, *et al.*, 2012).

Raciocínio causal (4) – Definições baseadas em raciocínio causal assumem um dado grafo causal[58]. Nele, as relações entre atributos e sua influência no resultado são capturadas por um conjunto de equações estruturais que são usadas para fornecer métodos para estimar efeitos de atributos sensíveis e criar algoritmos que garantam um nível tolerável de discriminação devido a esses atributos.

Justiça contrafactual. Um grafo causal é contrafactualmente justo se o resultado previsto no grafo não depender de um descendente do atributo protegido (KUSNER, LOFTUS, *et al.*, 2017).

---

[58] Grafo direcionado, acíclico, com nós que representam atributos de um candidato e arestas que representam relacionamentos entre os atributos.

Discriminação não resolvida. Um grafo causal não apresenta discriminação não resolvida se não houver um caminho do atributo protegido para o resultado previsto, exceto por meio de uma variável de resolução (KILBERTUS, ROJAS-CARULLA, *et al.*, 2017).

Discriminação via proxy[59]. Um grafo causal está livre de discriminação via proxy se não existir um caminho do atributo protegido para o resultado previsto que seja bloqueado por uma variável de proxy (KILBERTUS, ROJAS-CARULLA, *et al.*, 2017).

Inferência justa. Essa definição classifica os caminhos em um grafo causal como legítimos ou ilegítimos (NABI e SHPITSER, 2018).

Todas essas métricas estatísticas até aqui vistas, serão empregadas nos algoritmos descritos na seção 7.3.2.3 Algoritmos de mitigação e explicados na seção A.3 Imparcialidade, de acordo com os critérios que veremos a seguir.

7.3.2.2.1 População: Individual vs. Grupo

Justiça individual, lato sensu, busca que indivíduos semelhantes sejam tratados da mesma forma. Justiça de grupo, lato sensu, divide uma população em grupos definidos por atributos protegidos, i.e., raça, cor, orientação sexual etc., e procura que alguma medida estatística seja igual entre os grupos.

Um algoritmo é dito consistente ou individualmente justo se para pessoas semelhantes são geradas previsões semelhantes. Geralmente a medição desse atributo é feita por uma pontuação de consistência, i.e., uma estimativa precisa dos vizinhos mais próximos de cada ponto analisado. Observe que pessoas semelhantes a um indivíduo em um grupo protegido também podem pertencer a esse grupo, e talvez todas elas sejam tratadas igualmente mal.

---

[59] Um atributo proxy é um atributo cujo valor pode ser usado para derivar o valor de outro atributo

Já a justiça de grupo é mensurada por proporções, i.e., se a fração de pessoas com deficiência que não recebe um benefício é muito maior do que a fração de pessoas sem deficiência que não recebem um benefício, então, as pessoas com deficiência podem alegar que estão sendo discriminadas (PEDRESCHI, RUGGIERI e TURINI, 2012).

Outras medidas usadas para comparar o grupo protegido com o grupo não protegido incluem: diferença de risco (citada na lei britânica[60]), razão de risco ou risco relativo (citada na Corte de Justiça Europeia[61]), chance relativa e razão de probabilidade. E para comparar o grupo protegido com a população em geral temos: diferença de risco estendida; taxa de risco estendida ou elevação estendida; e chance estendida (preferida pelas cortes americanas[62]) (PEDRESCHI, RUGGIERI e TURINI, 2012).

Outras medidas de discriminação incluem: diferença de médias; diferença de coeficientes de regressão; testes de classificação; informação mútua (entre previsão e atributo protegido); diferença inexplicada (resíduos de previsões construídas com atributos não protegidos); e consistência (comparação de previsão com vizinhos mais próximos) (ŽLIOBAITė, 2015).

7.3.2.2.2 Fase de atuação: Dados vs. Modelo

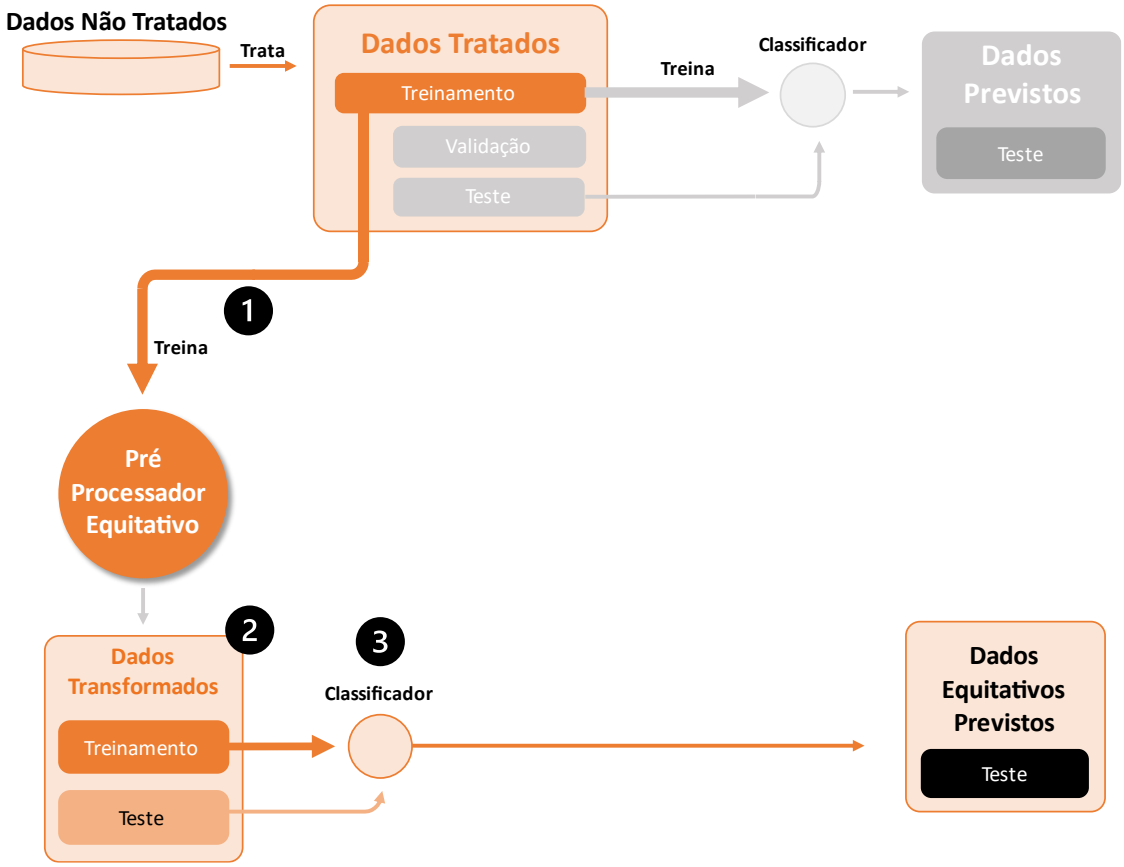
A neutralidade pode ser medida nas diferentes fases de confecção da IA (tanto nos dados de treinamento quanto no modelo). No modelo, podemos mensurar nas fases de pré-treinamento, in-treinamento e pós-treinamento (D'ALESSANDRO, O'NEIL e LAGATTA, 2017).

> **Pré-treinamento** – atacam o problema removendo a discriminação subjacente aos dados antes de qualquer modelagem (D'ALESSANDRO, O'NEIL e LAGATTA, 2017). ❶Um pré-processador é treinado com os

[60] Vide (REINO UNIDO, 1975)
[61] Vide (SCHIEK, WADDINGTON e BELL, 2007)
[62] Vide (ESTADOS UNIDOS DA AMÉRICA, 1976), (ESTADOS UNIDOS DA AMÉRICA, 1968), (ESTADOS UNIDOS DA AMÉRICA, 1967)

dados de treinamento originais e; ❷gera um novo conjunto de dados (dados transformados); ❸que é usado para treinar a IA que gera previsões justas (ilustrado na Figura 8)

Figura 8 – Esquema do funcionamento dos algoritmos de pré-processamento.



Fonte: d'Alessandro et al. (2017)

In-treinamento – As técnicas de In-treinamento podem ser consideradas modificações dos algoritmos tradicionais de aprendizado para lidar com a discriminação durante a fase de treinamento do modelo (D'ALESSANDRO, O'NEIL e LAGATTA, 2017). Abaixo (ilustrado na Figura 9) um exemplo de in-treinamento implementado por *generative adversarial network* (GAN).

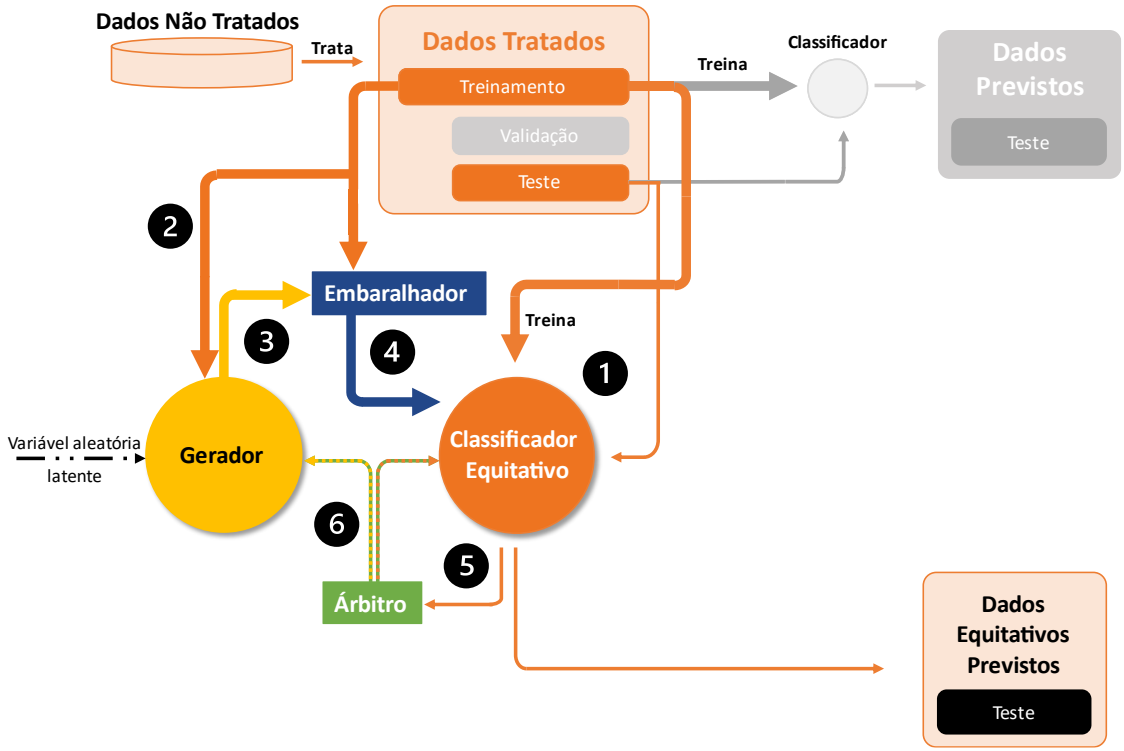
A GAN é composta por duas redes neurais, rede generativa e rede discriminativa, que competem entre si em um jogo. A rede generativa gera candidatos enquanto a rede discriminativa os avalia. O objetivo da rede generativa é aumentar a taxa de erro da rede discriminativa, i.e., nesse caso gerar previsões

que enganem o detector de neutralidade, o que na prática implicaria em um classificador justo.

❶A GAN é treinada usando dados reais e imparciais (parte do conjunto de dados de treinamento). Isso configura o modelo discriminador para distinguir entre dados enviesados e imparciais. ❷Produz a distribuição de dados que o gerador usará para produzir dados enviesados.

❸O gerador pega um vetor de dados aleatórios e os transforma com base na distribuição gaussiana, retornando um dado enviesado. ❹Essas entradas geradas, juntamente com algumas entradas imparciais do conjunto de dados de treinamento, são alimentadas no modelo discriminador (via embaralhador). ❺O Classificador Equitativo renderizará uma previsão probabilística sobre a natureza da entrada recebida, gerando um valor entre 0 e 1, onde 1 é tipicamente dado imparcial e 0 é um dado enviesado.

Figura 9 – Esquema do funcionamento de um algoritmo GAN.

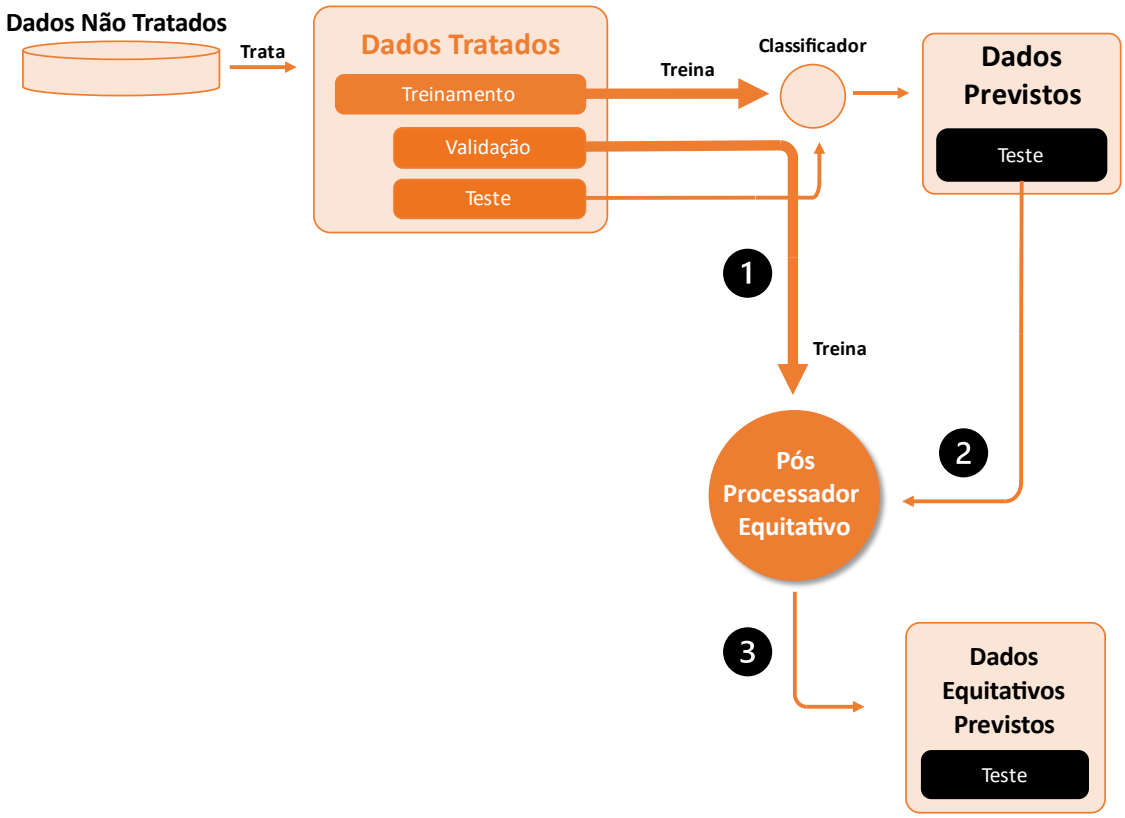


Fonte: d'Alessandro et al. (2017)

Os modelos Gerador e Classificador Equitativo estão jogando um jogo de soma zero um com o outro. ❻Quando o Classificador Equitativo é capaz de distinguir com sucesso entre exemplos imparciais e enviesados, nenhuma alteração é feita nos seus parâmetros. No entanto, grandes atualizações são feitas nos parâmetros do modelo quando ele não consegue distinguir entre dados probos e tendenciosos. O inverso é verdadeiro para o modelo Gerador, ele é penalizado (e seus parâmetros atualizados) quando falha em enganar o modelo Classificador Equitativo, mas caso contrário seus parâmetros permanecem inalterados (ou são recompensados).

Pós-treinamento – é a classe final de métodos e pode ser realizada após o treinamento. Ele se baseia no acesso a um conjunto de dados (validação) que não estava envolvido na fase de treinamento do modelo (D'ALESSANDRO, O'NEIL e LAGATTA, 2017).

Figura 10 – Esquema do funcionamento dos algoritmos de pós-processamento.



Fonte: d'Alessandro et al. (2017)

Os algoritmos de pós-treinamento, modificam as previsões. ❶ Um pós-processador é treinado com os dados de validação originais. ❷ Em seguida, ele recebe as previsões do modelo original como entrada e ❸gera um novo conjunto de previsões (previsões justas) (ilustrado na Figura 10).

7.3.2.2.3 Visão: Equidade vs. Isonomia

Existem duas visões de mundo opostas sobre justiça de grupo: “igualdade de resultados”, i.e., equidade, e “igualdade de tratamento", i.e., isonomia (YEOM e TSCHANTZ, 2018). A visão de mundo equitativa sustenta que todos os grupos tenham habilidades semelhantes em relação à tarefa (mesmo que não possamos observar isso adequadamente), enquanto a visão de mundo isonômica sustenta que as observações refletem a capacidade com relação à tarefa (FRIEDLER, SCHEIDEGGER e VENKATASUBRAMANIAN, 2016).

Por exemplo, a admissão no ensino superior, usa a nota do Exame Nacional do Ensino Médio (ENEM) como um recurso para prever o sucesso do candidato na faculdade. Segundo a visão isonômica, a pontuação se correlaciona bem com o futuro sucesso do candidato e existe uma maneira de usar o ENEM para comparar corretamente as habilidades dos candidatos. Por outro lado, a visão de mundo equitativa nos diz que a pontuação do ENEM pode conter vieses estruturais; portanto, sua distribuição sendo diferente entre os grupos, não deve ser confundida com uma diferença na distribuição da capacidade.

7.3.2.3 Algoritmos de mitigação

Os algoritmos de mitigação de viés tentam melhorar as métricas de justiça, modificando os dados de treinamento, o algoritmo de aprendizado ou as previsões. Essas categorias de algoritmos são conhecidas como pré-

treinamento, in-treinamento e pós-treinamento, respectivamente (D'ALESSANDRO, O'NEIL e LAGATTA, 2017).

As principais técnicas de mitigação de viés são: Pré-processamento Otimizado, Pós-processamento de Probabilidades Equalizadas, Repesagem, Regularizador Removedor de Preconceitos, Gerador de Representações Justas e Adversarial Debiasing. Essas técnicas são apresentadas de forma resumida no item A.3 Imparcialidade do anexo 1.

7.3.2.4 Limites

A essa altura o leitor deve estar se indagando sobre a factibilidade de agruparmos todas essas técnicas numa bateria de testes e analisarmos todas as métricas, vistas no item 7.3.2.2 Métricas, com o intuito de eliminarmos toda e qualquer possibilidade de viés em nossa IA. Infelizmente, como veremos a seguir, ao adotarmos uma das duas visões, i.e., equidade ou isonomia, anulamos a possibilidade de, concomitantemente, termos justiça por grupo e justiça individual.

Para (FRIEDLER, SCHEIDEGGER e VENKATASUBRAMANIAN, 2016) estudar justiça algorítmica é estudar as interações entre os diferentes espaços que compõem a cadeia de deliberação da tarefa. Uma tarefa é um mapeamento entre esses espaços, especificamente, entre o espaço construto e o espaço decisório ou entre o espaço observado e o espaço decisório.

O Espaço Construto (EC) é o espaço métrico que contém os atributos que queremos usar em nossas decisões. Na realidade, podemos não conhecer esses atributos ou mesmo a verdadeira semelhança entre os indivíduos. Daí a necessidade do espaço observado. O Espaço Observado (EO) é o espaço métrico que contém tudo o que podemos medir e observar. Podemos definir um mapeamento, denominado observação, que leva uma pessoa P de EC para uma pessoa P' de EO. O Espaço Decisório (ED) é o espaço métrico que contém as possíveis decisões.

Como todos os três espaços são métricos, podemos quantificar as transformações entre eles, i.e., medir a qualidade dessas transformações entre espaços em termos de distância. Para, tal são definidas 7 distâncias: distorção aditiva, medida de par, distância de Wasserstein (DW), distância de Gromov-Wasserstein (DGW), distância entre grupos (DEG), distância intragrupo (DIG) e viés de grupo (VG).

O axioma da equidade nos diz que todos os grupos são essencialmente o mesmo. A equidade, portanto, é uma propriedade do EC. Logo, não existem diferenças inatas entre grupos de indivíduos no EC. Assim, se a decisão tiver um caráter equiparativo, deve-se assumir o axioma da equidade.

O axioma da isonomia nos diz que: Existe um mapeamento entre EC e EO para o qual a distorção é mínima. Isso quer dizer que, se a decisão estiver sendo tomada considerando-se apenas os esforços de um indivíduo, o axioma da isonomia pode ser a escolha certa.

Podemos definir justiça como: Um mapeamento entre EC e ED é dito justo se objetos que são próximos em EC também são próximos em ED.

Podemos definir viés estrutural como a existência de mais DEG que DIG quando mapeamos de EC para EO, i.e., quando grupos são tratados diferencialmente pelo processo de observação. Repare que nossa definição não leva em consideração a equidade. Isso quer dizer que podemos ter, concomitantemente, viés estrutural e diferenças entre grupos agindo no EC separando os grupos no EO.

Podemos definir Discriminação Direta como como a existência de mais DEG que DIG quando mapeamos de EO para ED, i.e., quando grupos são tratados diferencialmente pelo processo de decisão.

Podemos definir Não Discriminação como evitar concomitantemente viés estrutural e discriminação direta. Ou, mais formalmente, como quando o mapeamento entre EC e ED possui um VG menor que um certo parâmetro. O objetivo da não discriminação é garantir que o processo de tomada de decisão não varie com base na participação no grupo.

No intuito de concluir nosso estudo sobre os mapeamentos, definimos um mecanismo como um mapeamento não trivial entre EO e o ED. Um mecanismo é dito rico se é sobrejetivo.

Podemos definir Mecanismo Individualmente Justo (MIJ) como o mecanismo de tomada de decisão que trata as pessoas de maneira semelhante se elas estiverem próximas e pode tratá-las de forma diferente se estiverem distantes, no EO.

Podemos definir Mecanismo Socialmente Justo (MSJ) como o mecanismo de decisão que trata todos os grupos da mesma maneira.

De posse dessas definições, podemos concluir que: sob o axioma da isonomia um MIJ garante a justiça, à exceção de um pequeno erro; um ED discreto não comporta um mecanismo justo e impede a justiça; e sob o axioma da equidade um MSJ garante a não discriminação. Vale destacar que a não discriminação se mantém mesmo sob a presença de viés estrutural, i.e., não é feita nenhuma suposição sobre o mapeamento do EC para o EO.

Sob a égide da isonomia, pode-se garantir justiça, enquanto sob uma visão de mundo equitativa, a não discriminação pode ser garantida. Aqui cabe o seguinte questionamento: essas visões de mundo são fundamentalmente conflitantes ou existem mecanismos que podem garantir justiça ou não discriminação nas duas visões de mundo?

Infelizmente, a isonomia parece ser crucial para garantir a justiça, i.e., se existe um viés estrutural no processo de decisão, nenhum mecanismo pode garantir a justiça. A justiça só pode ser alcançada, sob a visão de mundo isonômica, usando um mecanismo de justiça individual, e o uso de um mecanismo de justiça de grupo será injusto nessa visão de mundo.

E a não discriminação? Infelizmente, um simples contraexemplo mostra novamente que esses mecanismos não são agnósticos à visão de mundo. Embora os mecanismos de justiça de grupo demonstrem sua eficácia em prover a não discriminação, assumindo o princípio da equidade e do viés estrutural, se o viés estrutural for assumido, aplicar um mecanismo de justiça individual causa-

rá discriminação do espaço de decisão, que o princípio da equidade seja assumido ou não.

Fica demonstrado então que algumas noções de justiça são fundamentalmente incompatíveis entre si. Os resultados mostram que não é possível progredir sem uma definição precisa das crenças, i.e., sobre a visão de mundo, e dos tipos de danos que se deseja evitar.

Daí, ser crucial a escolha coerente das métricas e algoritmos de mitigação condizentes com o que se espera da IA, e sua relação com o mundo, e não um simples emprego ad hoc das técnicas. Face o exposto, cabe aqui uma perquisição: poderíamos usar apenas métricas favoráveis para explicar nossa IA de maneira a fazê-la parecer justa, i.e., promover um banho de isenção?

7.3.2.5 Problemas

Dieselgate é um nome cunhado pela imprensa para o escândalo de testes de emissão de poluentes envolvendo vários fabricantes de automóveis em todo o mundo (CHAPPELL, 2015). A Agência de Proteção Ambiental dos EUA descobriu um software instalado nos veículos da Volkswagen que alterava os números de emissão de poluentes somente quando os carros eram testados. As investigações subsequentes descobriram a mesma prática em outros países e por outros fabricantes, levando a uma das maiores crises da história da indústria automobilística. Agora imagine o emprego de algumas das mesmas técnicas de neutralidade, discutidas anteriormente, por vários fabricantes de IA com o propósito de mascarar comportamentos dúbios ou discriminatórios.

Fairwashing (banho de isenção) é a prática de promover uma falsa percepção de que um modelo de aprendizado de máquina respeita alguns valores éticos. Devido à crescente importância dos conceitos de justiça no aprendizado de máquina – e como o direito à explicação na GDPR não fornece diretrizes precisas sobre o que significa fornecer uma “explicação válida” – existe uma brecha legal que pode ser explorada por empresas desonestas para encobrir

alguns comportamentos, tidos como injustos, de seus modelos de caixa preta, fornecendo explicações equivocadas (i.e., racionalização).

Uma racionalização consiste em encontrar um modelo substituto interpretável (S) que explique um modelo de caixa preta (P), de modo que S seja mais justo que P. Para demonstrar a isenção de P, o modelo substituto S, obtido por meio da racionalização, pode ser mostrado ao auditor (e.g., uma entidade externa dedicada ou os próprios usuários) para convencê-lo de que a empresa, e a IA, é "limpa". É por isso que as auditorias das agências reguladoras, testes independentes e um ciclo de certificação robusto são vitais para mitigar essa prática.

Algumas das técnicas que vimos até aqui fornecem explicações, métricas e soluções para problemas de vieses encontrados, e outras, para prover inspeção e explicação sobre o modelo. O LaundryML tenta resolver alguns desses problemas; mas, por enquanto, nenhum algoritmo pode estimar se uma explicação é uma racionalização ou é capaz de detectar um “banho de isenção” (AÏVODJI, ARAI, *et al.*, 2019). Há, portanto, a necessidade de um ser humano na análise e detecção dessas tentativas de escamotear as “intenções” da IA.

Estes dois grupos de técnicas, i.e., explicabilidade e justiça, podem ser compilados em baterias de testes a serem aplicados à IA. Deles, podemos entender, numa visão global, local ou mesmo uma previsão específica, como a IA funciona e se ela apresenta, ou não, algum comportamento discriminatório, e o porquê. Esse nível de entendimento é vital para que possamos explicar a engenheiros, tomadores de opinião, usuários e autoridades desde o mero funcionamento, em linhas gerais, da IA até embasar uma argumentação jurídica.

Neste capítulo aprendemos sobre o escopo das divulgações a serem feitas (7.1 Quais divulgações serão feitas), os tipos de informações a serem coletadas (7.2 Quais tipos de informações serão coletados) e como obtê-las via XAI (7.3 XAI). Começamos por apresentar os cômputos internos (7.1.1 Cômputos internos) – i.e., aqueles que fazemos para diretores, programadores etc. – e externos (7.1.2 Cômputos externos) – i.e., aqueles que fazemos para investidores, consumidores, judiciário etc. – caracterizá-los e diferenciá-los. Em seguida,

analisamos os recursos manuseados para confecção dos proclames e expomos o que medir (7.2.1 O que medir), onde medir (7.2.2 Onde medir) – i.e., cadeia de suprimentos (7.2.2.1 Cadeia de suprimentos) e análise do ciclo de vida (7.2.2.2 Análise do Ciclo de Vida (ACV)) – e como medir (7.2.3 Como medir).

Por fim, investigamos o arcabouço equivalente aplicável a IA. Nele expomos o conceito de explicabilidade (7.3.1 Explicabilidade), uma taxonomia dos métodos (7.3.1.1 Taxonomia dos Métodos de Explicabilidade), algumas das principais técnicas (7.3.1.2 Técnicas de Explicabilidade) e um subconjunto sobre imparcialidade (7.3.2 Imparcialidade).

Saneadas as questões sobre quais tipos de informações coletar e quais divulgações fazer, no capítulo seguinte, nos debruçaremos sobre como divulgar as informações obtidas pela aplicação das diversas técnicas de IAE ao longo do ciclo de vida da IA para os diversos stakeholders.

8 COMO DIVULGAR AS INFORMAÇÕES

Todo resultado traz alguma forma de consequência, tanto para o demandado quanto para o demandante. As consequências aparecem quando um plácito é cristalino e seu cumprimento ou descumprimento são claramente visíveis (i.e., os resultados respondem sim ou não a pergunta “Você cumpriu sua promessa?”).

Ter um pacto definido com clareza e com um sistema de acompanhamento e mensuração desde o início é essencial se você pretende determinar suas consequências. Essas consequências podem ser positivas ou negativas. Contudo, mesmo quando negativas, é possível usá-las como uma oportunidade de aprendizado e amadurecimento.

Como versado ao longo dessa obra, ao instado, coube realizar certas ações de acordo com as expectativas pactuadas e fornecer uma avaliação dessas ações para os pleiteantes. O requisitante, acompanhou o progresso e manteve-se estoico quanto ao pactuado.

À vista disso, nos ocuparemos de responder ao longo desse capítulo a (QN4) “Como divulgar as informações?”. Prima facie quanto mais informados os diferentes stakeholders tiverem sobre a IA, maior sua capacidade de fazer uma escolha informada. Porém, como veremos no item 8.1 Pirâmide DIKW, o conhecimento existe ao longo de um continuum e não em coleções de dados e, portanto, devemos trabalhar nossos cômputos para melhor servir os receptores.

A maneira como iremos (e por vezes poderemos) conformar nossas explanações está atrelada principalmente a dois fatores: escopo (interno ou externo) e incidente (ex-ante ou ex-post). É claro que outros fatores incidiram sobre a formatação, e.g., natureza, finalidade etc., e havendo padronização disponível deve-se obtemperar, para que possamos garantir o máximo de atributos de uma boa informação (i.e., relevância, confiabilidade, comparabilidade, verificabilidade, compreensibilidade e disponibilidade).

Para atendermos nossas duas obrigações de *accountability* produzimos cômputos internos que embasam nossa tomada de decisão e realização de certas

ações. Igualmente, produzimos os cômputos externos para fornecer avaliação de nossas ações para os stakeholders.

Todas estas evidenciações se destinam a uma fase ex-ante. Para uma análise ex-post precisamos conduzir uma investigação do incidente. O processo usitado deve examinar todos os fatores subjacentes em uma cadeia de eventos que culminaram no incidente e identificar as causas do problema.

As divulgações internas ex-ante abarcam os cômputos internos que embasam nossa tomada de decisão e realização de certas ações. Por possuírem variegada finalidade, devem condizer com as necessidades da empresa (e.g., relatórios gerenciais, avaliação de desempenho etc.).

As divulgações externas ex-ante abarcam os cômputos externos que embasam tomada de decisão e realização de certas ações por parte dos stakeholders. Por possuírem finalidade específica, costumam ser regulamentadas (e.g., relatórios financeiros etc.).

As divulgações ex-post possuem finalidade específica, i.e., geralmente relatar erros[63], defeitos[64] e falhas[65] (e.g., relatórios de incidentes, relatórios de segurança etc.). Estes artefatos costumam ser o resultado de um processo investigativo que busca pela origem do fato.

Examinaremos no item 8.2 Análise de Acidente a técnica de Análise de Acidente e suas quatro etapas. Em seguida, no tópico 8.3 Análise da Causa Raiz (ACR), veremos as sete etapas da Análise da Causa Raiz (ACR) e como ela, juntamente com a Forense Digital (FD), item 8.4 Forense Digital, promoverão a preservação, coleta, validação, identificação, análise, interpretação, documentação e apresentação de evidências digitais derivadas de fontes digitais usadas para identificar as causas raiz de falhas ou defeitos.

---

[63] "A diferença entre um valor ou condição computada, observada ou medida e o valor ou condição verdadeira, especificada ou teoricamente correta."; "uma etapa, processo ou definição de dados incorreto."; "um resultado incorreto." ou "uma ação humana que produz um resultado incorreto." (IEEE, 1990).

[64] “Uma imperfeição ou deficiência em um produto de trabalho em que o produto de trabalho não atende a seus requisitos ou especificações e precisa ser consertado ou substituído.” (IEEE, 2009).

[65] “Rescisão da capacidade de um produto executar uma função exigida ou sua incapacidade de executar dentro dos limites especificados anteriormente” (ISO, 2005) ou “um evento no qual um sistema ou componente do sistema não executa uma função necessária dentro dos limites especificados” (ISO, 2009).

Por fim, no ponto 8.5, examinaremos o uso da XAI como técnica forense, i.e., Inteligência Artificial Explicativa Forense (IAEF). Ademais, explanaremos sobre as principais fases que diferem das nove etapas do processamento de evidências digitais nos tópicos: 8.5.1 Exame, 8.5.2 Análise e 8.5.3 Apresentação.

Antes de aplicarmos todos esses métodos inquisitivos, precisaremos coletar e evoluir os dados que embasarão essas divulgações. Começaremos por compreender como se dá o processo de apuração dos dados até se tornarem conhecimento.

8.1 PIRÂMIDE DIKW

Em 1934, T.S. Eliot escreveu:

"Onde está a sabedoria que perdemos no conhecimento?

Onde está o conhecimento que perdemos na informação?"

(ELIOT, 1934)

A hierarquia *data–information–knowledge–wisdom* (DIKW), ou dados-informação-conhecimento-sabedoria (DICS) em português, também referida como 'Hierarquia do Conhecimento', 'Hierarquia da Informação' ou 'Pirâmide do Conhecimento' é uma das estruturas fundamentais e amplamente reconhecidas nas literaturas de gestão da informação, sistemas de informação e gestão do conhecimento. Apesar de Eliot ser tido como o primeiro a descrevê-la, coube a (ACKOFF, 1989) em seu artigo *From data to wisdom*, 'Dos dados à sabedoria' em português, sua primeira descrição sistemática[66].

[66] A hierarquia proposta continha os seguintes níveis: data, information, knowledge, understanding and wisdom (dados, informação, conhecimento, compreensão e sabedoria em português).

A jerarquia é usada para contextualizar os elementos, em relação uns aos outros, e para identificar e descrever os processos envolvidos na transformação de uma entidade em um nível inferior (e.g., dados) para uma entidade em um nível superior na hierarquia (e.g., informação). A suposição implícita é que os dados podem ser usados para criar informações; as informações podem ser usitadas para criar conhecimento, e o conhecimento pode produzir sabedoria.

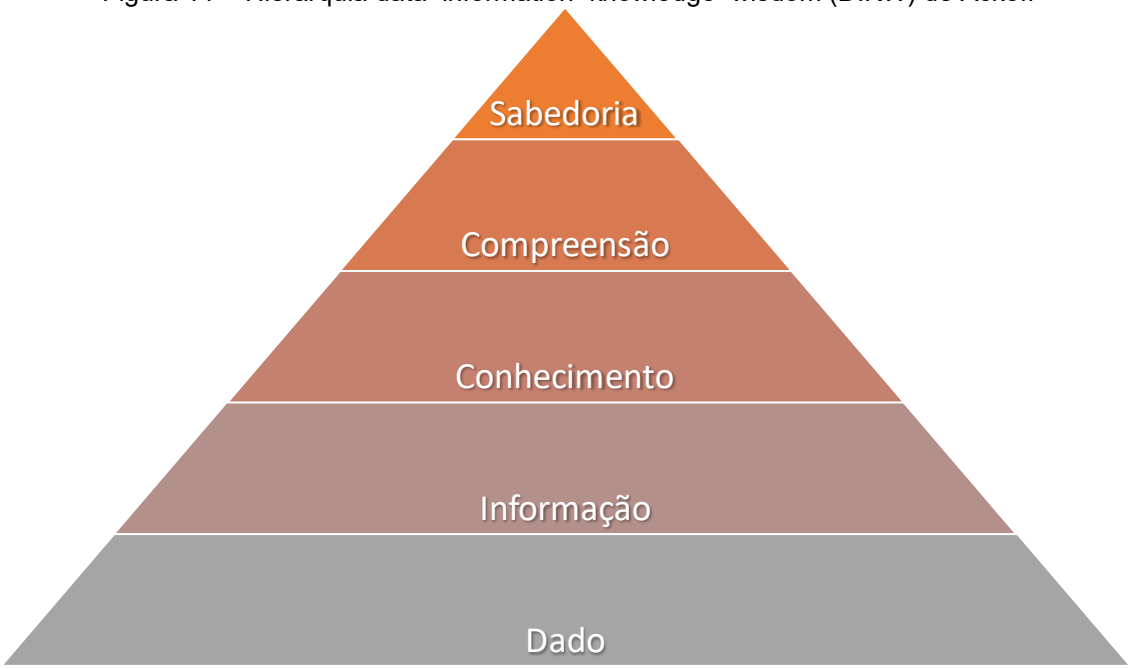
A Figura 11 – Hierarquia data–information–knowledge–wisdom (DIKW) de Ackoff ilustra a Pirâmide do Conhecimento proposta por Ackoff. Para o autor os cinco elementos, dado, informação, conhecimento, compreensão e sabedoria, são definidos da seguinte forma:

> **Dados** – símbolos. Os símbolos representam propriedades de objetos, eventos e seu ambiente. Eles são os produtos da observação. Mas são inúteis até que estejam em uma forma utilizável (i.e., relevante). A diferença entre dados e informações é funcional, não estrutural.
>
> **Informação** – A informação é inferida a partir dos dados (i.e., dados são processados para serem úteis). A informação está contida em descrições e respostas a perguntas como: ‘quem’, ‘o quê’, ‘quando’ e ‘quantos’.
>
> **Conhecimento** – aplicação de dados e informações. Conhecimento é know-how (i.e., saber como), e é o que possibilita a transformação de informações em instruções. O conhecimento pode ser obtido por transmissão (de outro que o possui), por instrução ou experimentação.

“O conhecimento existe ao longo de um continuum entre o conhecimento tácito (saber como) e o conhecimento explícito (saber o quê)” (JASHAPARA, 2005). O conhecimento explícito pode ser registrado em sistemas de informação, enquanto o conhecimento tácito não pode ser registrado, pois faz parte da mente humana.

> **Compreensão** – apreciação do "porquê".
>
> **Sabedoria** – compreensão avaliada. É a capacidade de aumentar a eficácia. A sabedoria agrega valor, o que requer a função mental que chamamos de julgamento. Os valores éticos e estéticos que isso implica são inerentes ao ator e são únicos e pessoais.

Figura 11 – Hierarquia data–information–knowledge–wisdom (DIKW) de Ackoff



Fonte: Ackoff (1989)

Para Ackoff, as primeiras quatro categorias se relacionam com o passado (i.e., elas lidam com o que foi ou o que é sabido). Apenas a quinta categoria (sabedoria) lida com o futuro porque incorpora a capacidade de pôr em prática o comportamento mais adequado, tendo em conta o que se sabe (conhecimento e compreensão) e o que é benéfico (considerações éticas e sociais). Com sabedoria, as pessoas podem criar o futuro em vez de apenas compreender o presente e o passado.

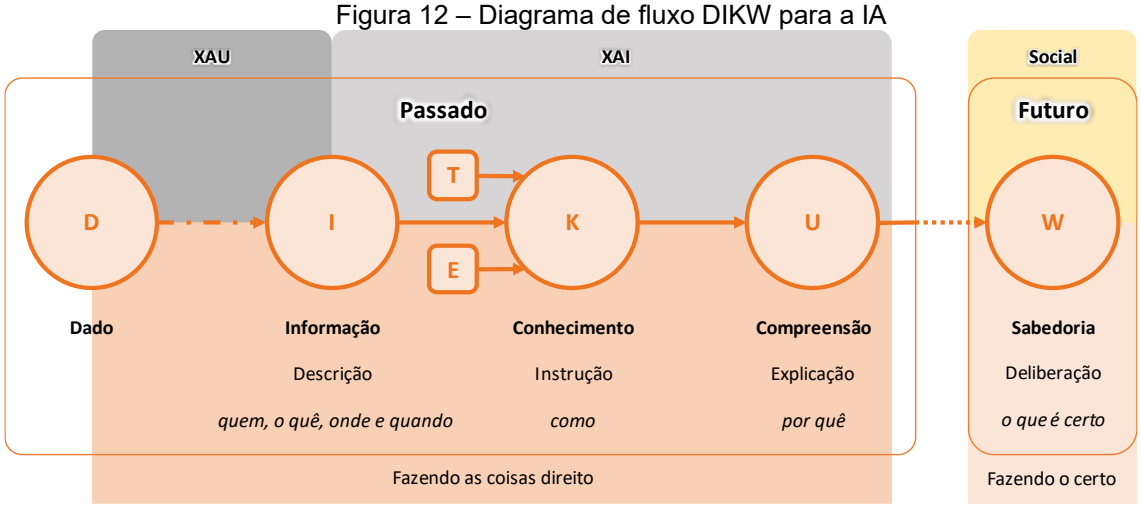
Mas alcançar a sabedoria não é fácil e, geralmente, as pessoas devem passar sucessivamente pelas outras categorias para embasarem sua procedência. Ou seja, dados reais e de qualidade geram informações (descrições e respostas a perguntas como: ‘quem’, ‘o quê’, ‘quando’ e ‘quantos’) fiáveis que por sua vez embasam a constituição de conhecimentos tácitos (saber como) e explícitos (saber o quê), culminando na apreciação do "porquê".

Por exemplo, uma negativa de empréstimo para um determinado cliente do banco pela IA num determinado dia é um **dado**. Sozinha, ela não diz nada além de um fato objetivo. Contudo, ao analisar um conjunto de negativas ao longo de um certo período (e.g., uma semana, um mês etc.), é possível avaliar se ocorreu viés, ou seja, cria-se **informação**. Ao pensar em *como* isso pode representar cenários positi-

vos ou, então, situações de crise (e.g., isso afetará a imagem do banco), há o **conhecimento** sobre a situação. A partir da **compreensão** e diagnóstico do cenário viabilizamos as respostas necessárias para uma deliberação e eventual criação de um plano de ação para esse fim, i.e., a **sabedoria** sendo implementada.

A Figura 12 – Diagrama de fluxo DIKW para a IA – ilustra como aplicar as respostas das questões normativas (QN1), (QN2) e (QN3) para responder a (QN4). Essencialmente o que queremos com a QN4 é propiciar aos stakeholders (QN2), solicitantes dos cômputos sobre a IA, meios (QN3) para avaliarem o cumprimento do pacto (QN1).

Figura 12 – Diagrama de fluxo DIKW para a IA



Fonte: O autor (2023)

O fluxograma acima (Figura 12 – Diagrama de fluxo DIKW para a IA) fornece uma visão geral da transformação dos dados sobre a IA. Ela contém as fases do DIKW (○), suas respectivas denominações (**negrito**) e os tipos de perguntas a que se propõem responder cada fase (*itálico*).

O nosso sistema de telemetria (XAU) abarca os estágios de Dado e Informação. A XAU deve levar em conta não apenas o entendimento e a preparação dos dados, mas também o processo, os objetivos, os requisitos e as pessoas designadas na fase de entendimento do negócio, estabelecendo assim uma linhagem de dados e, portanto, ela começa a coletar dados desde a fase de compreensão do negócio. Como já mencionado, o sistema de telemetria é ferramenta essencial na fase

de implantação para monitorar o produto durante seu ciclo de vida e como um registrador de dados.

Os cômputos produzidos pelos petrechos de XAI cobrem os estágios de Informação, Conhecimento e Compreensão. Quase todas as nossas técnicas de XAI são empregadas principalmente, nas fases de modelagem e avaliação do CRISP-DM para fins de depuração (ou seja, ajustes e correções), conformidade, inteligibilidade, imparcialidade e outras metas definidas na fase de entendimento do negócio. Podemos adaptar essas técnicas para fins ad hoc, por exemplo, imparcialidade ou robustez, o que pode nos levar a adotar etapas anteriores do processo, por exemplo, compreensão e preparação de dados, para sanar quaisquer inconsistências encontradas.

Em suma, nos valeremos dos dados e informações coletados pela XAU como insumos para o multifário arsenal de XAI que engendrará uma série de descrições (*quem*, *o quê*, *onde* e *quando*), conhecimentos (*como*) e compreensões (*porque*) sobre a IA. Daí, resta-nos apenas conformar essa apreensão para que as partes interessadas possam deliberar no estágio de Sabedoria (e.g., “A IA teve culpa no acidente?”, “Quais dados geraram o viés?”, “A IA agiu de forma ética?” ou “Como podemos evitar que isso ocorra novamente?”).

A maneira como iremos (e por vezes poderemos) conformar nossas explanações está atrelada principalmente a dois fatores: a natureza dos nossos stakeholders (i.e., internos ou externos) e o fato (i.e., ex-ante ou ex-post). É claro que outros fatores incidiram sobre a formatação, e.g., natureza, finalidade etc., e havendo padronização disponível deve-se sempre optar por, para que possamos garantir o máximo de atributos de uma boa informação (i.e., relevância, confiabilidade, comparabilidade, verificabilidade, compreensibilidade e disponibilidade).

Para atendermos nossas duas obrigações de *accountability* produzimos cômputos internos que embasam nossa tomada de decisão e realização de certas ações. Igualmente produzimos os cômputos externos para fornecer avaliação de nossas ações para os stakeholders.

Como explanado anteriormente, os cômputos internos são informações sobre a posição e o desempenho, da empresa ou da IA, destinados ao uso por pessoal

interno a organização, e.g., gerentes, diretores, funcionários etc., e variam em qualidade, conteúdo, formato, apresentação etc. Os cômputos externos são informações sobre a posição e o desempenho, da empresa ou da IA, destinados ao uso por pessoas externa a organização, e.g., acionistas, reguladores etc., e costumam ser regulamentados com relação aos procedimentos a serem usados para gerar as demonstrações de propósito geral.

Todavia, ainda há pouca regulamentação que rege os relatórios externos de desempenho da IA. A AAA22 menciona duas demonstrações: avaliação de impacto na proteção de dados e avaliação de impacto do sistema de decisão automatizada.

A avaliação de impacto na proteção de dados é um estudo que avalia até que ponto um sistema de informação protege a privacidade e a segurança das informações pessoais que o sistema processa. Idealmente, o ecossistema da IA (i.e., ciclo de vida, cadeia de fornecedores, XAU etc.) deve se adequar às normas da: família de padrões ISO/IEC 27000, que ajuda as organizações a manter seguros os ativos de informações; ISO 31000, para gerenciamento de riscos, sendo incorporada aos métodos de prevenção e mitigação de risco da empresa, caso adote; e ISO 19011:2018, que apresenta diretrizes para auditoria de sistemas de gerenciamento.

A avaliação de impacto do sistema de decisão automatizada é o estudo que avalia o sistema de decisão automatizada e o processo de desenvolvimento do sistema, incluindo design e dados de treinamento do modelo, quanto a impactos na precisão, justiça, preconceito, discriminação, privacidade e segurança[67]. Todas essas informações requeridas podem ser obtidas junto aos diversos artefatos que documentam o processo de confecção da IA[68], pela adoção das referidas normas ISO e emprego das técnicas e artefatos da XAU e da XAI.

Todas estas evidenciações se destinam a uma fase ex-ante. Para uma análise ex-post (e.g., para decidir quem será responsável em casos de perda ou dano resultante de uma deliberação da IA, para cumprir requisitos legais, determinar origens de falhas etc.) precisamos empregar algum processo inquiridor e não só modelos predefinidos de demonstrações. Devemos conduzir uma investigação do inciden-

---

[67] Vide ‘avaliação de impacto do sistema de decisão automatizada’ em 4.6 Políticas.
[68] Vide saídas em 6.2.1 CRoss-Industry Standard Process Model (CRISP-DM) e subitens.

te. O processo usitado deve examinar todos os fatores subjacentes em uma cadeia de eventos que culminaram no incidente e identificar as causas do problema.

8.2 ANÁLISE DE ACIDENTE

A metodologia que discutiremos a seguir é a mesma utilizada pelo NTSB nas investigações dos acidentes da Ilustração 1e Ilustração 2, e não difere daquela utilizada nas investigações em geral. A razão é que essa metodologia investigativa consagrada foi adaptada por setores específicos (e.g., aviação, energia nuclear, transporte ferroviário, medicina etc.) para atender a determinadas particularidades de cada um, sem nenhum prejuízo.

A investigação é necessária para examinar todos os fatores subjacentes em uma cadeia de eventos que culminaram em um acidente. Mesmo os incidentes aparentemente mais diretos, e.g., a Ilustração 2, raramente dependem de uma única causa. Por exemplo, na Ilustração 1, uma "investigação" que conclui que o ADS falhou em detectar as faixas na pista, e não vai além, não consegue encontrar respostas para várias questões importantes: “Por que o carro não detectou as faixas na pista?”, “Foi a sinalização da rua que falhou? Se sim, uma placa resolveria o problema?”, “Um carro equipado com LIDAR teria o mesmo problema?”, “Foi hackeado?” ou “A IA do carro não poderia identificar a faixa contrária de outra forma?”.

Uma parte crucial do processo de investigação de acidentes é a análise. Uma análise de acidente é realizada para determinar a causa (ou causas) de um acidente. Na Tabela 10 – Quatro etapas da análise de acidentes apresentamos uma versão resumida do processo de investigação que consiste em quatro etapas principais: coleta de fatos, análise de fatos, conclusões e contramedidas.

Os modelos de análise de acidente são agrupados em duas categorias principais: modelos de acidentes sequenciais e modelos de acidentes sistêmicos. Os modelos de acidentes sequenciais descrevem um acidente como uma cadeia de eventos discretos que ocorrem em uma ordem temporal específica. Eles empregam técnicas tradicionais, e.g., modos de falha e análise de efeitos (FMEA), análise de árvo-

re de falhas (FTA), análise de árvore de eventos e análise de causa-consequência. Geralmente são empregados em investigações de falhas de componentes físicos e erros humanos envolvendo sistemas relativamente simples.

Tabela 10 – Quatro etapas da análise de acidentes

	Etapa	Descrição
1	Coleta de fatos	Emprega um conjunto de processos forenses para preservar, coletar e documentar evidências para reunir todos os fatos possivelmente relevantes que possam contribuir para a compreensão do acidente.
2	Análise de fatos	Após o processo forense ter sido concluído ou parcialmente entregue, os fatos são reunidos para ilustrar a sequência de etapas que levaram ao acidente, verificando a consistência e a plausibilidade.
3	Conclusões	Se o histórico do acidente for suficientemente informativo, as conclusões são tiradas, apoiadas pelas evidências, sobre quais fatores estão ligados às circunstâncias e consequências do acidente. Assim, são identificadas as causas do acidente que precisam ser corrigidas.
4	Contramedidas	Ações corretivas ou recomendações são feitas para evitar a recorrência de um acidente semelhante.

Fonte: O Autor

Os modelos de acidentes sistêmicos descrevem um acidente como uma ocorrência que surge das interações entre os componentes do sistema, i.e., uma rede complexa e interconectada de eventos. Os três modelos principais são AcciMap[69], o Functional Resonance Analysis Method (FRAM)[70] e o System Theoretic Accident

[69] Vide (RASMUSSEN, 1997).
[70] Vide (HOLLNAGEL, 2012).

Model and Processes (STAMP)[71], geralmente empregados em casos de sistemas complexos porque incluem os princípios, modelos e leis necessários para entender inter-relações e interdependências complexas entre componentes (técnicos, humanos, organizacionais e gerenciais) de um sistema complexo.

Apesar de escolhermos um modelo específico – a saber, Análise da Causa Raiz (ACR) – para explorarmos na próxima seção, não há perda de generalidade. Modelos de acidentes sequenciais ou sistêmicos podem ser usados na análise de sistemas de IA sem prejuízo a adoção de técnicas de XAI.

8.3 ANÁLISE DA CAUSA RAIZ (ACR)

Em ciência e engenharia, ACR é o principal método empregado na análise de acidentes. É um método de resolução de problemas usado para identificar as causas raiz de falhas ou problemas (WILSON, DELL e ANDERSON, 1993).

Os diferentes campos que aplicam ACR nem sempre usam a mesma denominação, número de fases ou o mesmo conjunto de técnicas por fases, mas todos têm o mesmo objetivo: definir a causa raiz. Além disso, esse método pode ser usado ex-ante, agindo para a prevenção de incidentes e ex-post, para reagir e mitigar os efeitos causados pelos problemas. Na Tabela 11 – Sete etapas da análise da causa raiz, apresentamos uma versão do RCA composta por sete etapas:

Agora focamos nossa atenção no passo 2, ‘Investigar os fatores’, porque as principais contribuições da XAU estão aqui. Embora esta etapa seja a principal fonte de informação, seus resultados permeiam todo o processo de análise (etapas 3 a 5), mitigação (etapa 6), descrição (etapa 7) e podem impactar as deliberações iniciais da etapa 1.

Informações reunidas sobre ações e condições (baseadas em evidências) são fatos e outras informações (e.g., uma condição que deveria existir, mas não existe, são contrafactuais). Nosso foco principal de atenção é o que aconteceu (i.e.,

---

[71] Vide (ALTABBAKH, ALKAZIMI, *et al.*, 2014)

os fatos). Em seguida (etapas 3 a 6), ao aplicarmos XAI, voltamo-nos para como e por que determinadas condições e ações não terem sido atendidas, (i.e., contrafactuais, pois estas têm influência indireta no desfecho do problema) e como mitigar as causas e suas extensões.

Tabela 11 – Sete etapas da análise da causa raiz

	Etapa	Descrição
1	Escopo do problema	Estabelecer o que aconteceu, quando aconteceu, onde aconteceu e quem estava envolvido, i.e., uma definição clara do problema que você está investigando.
2	Investigar os fatores	Decidir quais informações coletar e quem entrevistar. Coletar evidências físicas, revisar processos e procedimentos, fotografias ou filmar a cena etc.
3	Reconstruir a história	Recriar o incidente mostrando um fluxo ou sequenciamento lógico. Desenvolver uma linha do tempo detalhada para mostrar claramente o que aconteceu quando.
4	Estabelecer fatores causais	Identificar condições, situações ou ações que desencadearam, permitiram ou influenciaram o incidente, i.e., estabelecer fatores causais.
5	Validar fatores subjacentes	Encontrar as causas raiz para cada um dos fatores contribuintes do incidente. A extensão das causas (físicas, humanas, IA etc.) deve ser determinada.
6	Planejar ações corretivas	Desenvolver uma ou mais ações corretivas para eliminar ou controlar cada causa e sua extensão.
7	Relatar aprendizados	Fornecer um relatório formal, permanente, auditável e defensável de suas descobertas.

Fonte: Wilson et al. (1993)

No final desta etapa, as informações compiladas nos ajudarão a reconstruir o incidente. Com base nos fatos coletados, estabelecemos a cadeia de ações, juntamente com os fatores que afetaram o desempenho do hardware, da IA e do humano.

Para problemas menores, simples ou diretos, as descobertas da etapa 1 podem ser tudo o que é necessário para estabelecer uma causa e recomendar alguma ação para resolvê-la. Para incidentes e condições adversas mais significativos, uma análise de causa mais profunda (esquadrinhando sistematicamente todos os cenários possíveis que podem ter produzido o problema) deve abordar as raízes físicas, humanas e técnicas do problema, e.g., ilustrações Ilustração 1eIlustração 2.

Ao investigar uma falha de equipamento, a primeira linha de investigação deve ser destinada a determinar qualquer problema de forma, ajuste ou função que precise ser corrigido. Em seguida, devem ser identificados os quatro possíveis mecanismos de degradação: força, ambiente reativo, tempo e temperatura (BLOCH, 2005). Posteriormente, são avaliadas as interfaces homem-máquina primárias, i.e., os humanos que influenciaram ou permitiram que os fenômenos físicos existissem, a fim de buscar a(s) raiz(es) humana(s) do incidente (Bloch, 2011). Neste ponto, podemos ter uma IA substituindo esse humano (humano fora do loop) ou entre a máquina e o humano (humano no loop).

Embora uma IA não seja uma pessoa, podemos analisar o que ela fez e como, de três ângulos diferentes: a tarefa que estava realizando, o potencial da IA para ter sucesso na tarefa e o processamento das informações do trabalho. Como a IA é propensa a cometer erros, os engenheiros criam barreiras ou defesas para garantir a segurança. Assim, nossa próxima linha de inspeção deve buscar barreiras projetadas (e.g., segurança cibernética e técnicas de IA robustas), defesas administrativas, defesas de supervisão e defesas culturais (MUSCHARA, 2007).

As raízes de um incidente envolvendo IA geralmente têm origens latentes mais profundas, e.g., na camada de negócios. Toda essa coleta de dados e análises posteriores não apenas mostrarão o que a IA fez que levou ao incidente, mas também as circunstâncias nas quais ela deliberou. Para esse processo de descoberta, precisamos do máximo de evidências possíveis, especialmente sobre a IA, que nos

permita tanto corroborar os fatos quanto investigar as ações da IA. Apresentamos a seguir uma maneira de realizar essa tarefa.

8.4 FORENSE DIGITAL

Como observado anteriormente, entrevistar uma testemunha é o equivalente humano a empregar algumas técnicas de XAI para "entrevistar" a AI. Por entrevista entendemos uma conversa estruturada na qual um entrevistador faz perguntas e o entrevistado dá respostas sobre o que é inquirido baseado no que experenciou.

A entrevista é uma técnica que nos ajuda a obter informações, ideias, experiências e compreensão por meio do diálogo com outras pessoas. Buscar o que os outros sabem nos ajuda a expandir nossa compreensão do que, como e (às vezes) por quê algo aconteceu. O principal objetivo da entrevista, aqui, é estabelecer o contexto no qual o incidente ocorreu (e.g., objetivos, foco, sequência de ações, conhecimento e consciência situacional[72]), focando nos fatos e buscando entender por que, e não apenas o quê. Juntamente com a análise de outras evidências obtidas, as entrevistas constituem a base da nossa próxima etapa na análise de acidentes, i.e., reconstruir a história.

Idealmente, o nível de comunicação humano-IA nos permitiria entrevistar a IA como Del Spooner interrogou Sonny[73] ou analisá-los como Rick Deckard, ao aplicar o teste Voigt-Kampff em Rachael[74]. No entanto, ainda estamos longe desse nível de interação. Por enquanto, podemos usar nosso conjunto de evidências digitais, a maioria da XAU, e algumas técnicas de XAI como ferramentas forenses para analisar a IA e cumprir nossos requisitos legais.

Em geral, analisar todos os dados que foram adquiridos, não apenas pela XAU, e avaliá-los fornece a evidência digital que a investigação precisa. Como em qualquer investigação, devemos identificar os dados que verificam as teorias exis-

---

[72] Percepção de elementos e eventos ambientais em relação a: tempo e espaço, compreensão de seu significado e projeção de seu status futuro (ENDSLEY, 1995).
[73] Robô NS-5 da USR em ‘I, Robot, de Asimov’.
[74] Replicante em ‘Do Androids Dream of Electric Sheep?’, de Philip K. Dick.

tentes (evidências incriminatórias), contradizem outras (evidências ilibatórias) ou mostram sinais de adulteração para ocultar dados (CARRIER, 2002).

O ramo da ciência forense responsável por essa análise é a perícia forense digital (FD), definida por Palmer como:

O uso de métodos cientificamente derivados e comprovados para a preservação, coleta, validação, identificação, análise, interpretação, documentação e apresentação de evidências digitais derivadas de fontes digitais com o objetivo de facilitar ou promover a reconstrução de eventos ou ajudar a antecipar ações não autorizadas, comprovadamente perturbadoras das operações planejadas.
(PALMER, 2001)

As diferentes áreas que aplicam a perícia digital nem sempre usam a mesma denominação, número de fases ou o mesmo conjunto de técnicas por fases. Além disso, o FD pode ser usada ex-ante, antecipando ações não autorizadas, e ex-post, para facilitar ou promover a reconstrução de eventos. Na Tabela 12 – Nove etapas do processamento de evidências apresentamos uma versão do processamento de evidências digitais composta por nove etapas.

O leitor deve observar a semelhança entre algumas etapas do processo descrito acima e outras descritas neste trabalho. Essas etapas compõem uma instância (para crimes que envolvem evidências digitais) das práticas atuais que coletam evidências físicas. Outra questão válida diz respeito à não inclusão da análise de provas forenses (físicas ou digitais) no processo de ACR ou da análise de acidentes como etapa.

O analista forense digital (AFD) trabalha dentro do sistema de justiça, fornecendo evidências importantes para investigações criminais (por exemplo, relatórios forenses). Os policiais costumavam fazer o trabalho de um AFD; porém, com a massiva adoção e evolução da tecnologia, profissionais especializados na área tornaram-se necessários. Além disso, esses profissionais costumam ser auxiliados por desenvolvedores, fabricantes, engenheiros e cientistas em casos específicos.

Tabela 12 – Nove etapas do processamento de evidências digitais

	Etapa	Descrição
1	Identificação	Reconhecer um incidente a partir de indicadores e determinar seu tipo. Isso não está explicitamente dentro do campo forense, mas é significativo porque afeta outras etapas.
2	Preparação	Elaboração de ferramentas, técnicas, mandados de busca e acompanhamento de autorizações e apoio à gestão.
3	Abordagem	Formular dinamicamente uma abordagem com base no impacto potencial sobre os espectadores e na tecnologia específica em questão. O objetivo da estratégia deve ser maximizar a coleta de evidências não adulteradas e, ao mesmo tempo, minimizar o impacto sobre a vítima.
4	Preservação	Isolar, proteger e preservar o estado das evidências físicas e digitais, inclusive impedindo que as pessoas usem o dispositivo digital ou permitindo que outros dispositivos eletromagnéticos sejam usados dentro de um raio afetado.
5	Coleta	Gravar a cena física e duplicar a evidência digital usando procedimentos padronizados e aceitos.
6	Exame	Pesquisa sistemática aprofundada de evidências relacionadas ao crime suspeito. Isso se concentra na identificação e localização de possíveis evidências, possivelmente em locais não convencionais e na construção de documentação detalhada para análise.
7	Análise	Determinar a significância, reconstruir fragmentos de dados e tirar conclusões com base nas evidências encontradas. Apoiar uma teoria do crime pode levar vá-

		rias iterações de exame e análise. A distinção da análise é que ela pode não exigir altas habilidades técnicas para ser executada e, portanto, mais pessoas podem trabalhar nesse caso.
8	Apresentação	Resumir e explicar as conclusões. Isso deve ser escrito em termos leigos usando terminologia abstrata. Toda terminologia abstrata deve fazer referência aos detalhes específicos.
9	Retorno	Garantir que a propriedade física e digital seja devolvida ao devido proprietário, bem como determinar como e quais evidências criminais devem ser removidas. Novamente, esta não é uma etapa forense explícita; no entanto, qualquer modelo que apreende evidências raramente aborda esse aspecto.

Fonte: Reith et al. (2002)

Os relatórios forenses servem como um documento que descreve as evidências, procedimentos e análises empregadas pelo AFD para apoiar suas conclusões. Esses relatórios servem como entrada para outros processos investigativos para reconstruir a história, estabelecer ou refutar fatores contribuintes, validar ou invalidar fatores subjacentes e ajudar a planejar ações corretivas.

A perícia pode ser solicitada pelos juízes ou apresentada pelas partes (autor e réu); portanto, é um processo independente. Por esse motivo, afirmamos anteriormente que o método de análise de acidentes empregado não afeta o uso da XAI, mas afeta a forma como o caso é investigado.

Além disso, como mencionado anteriormente, a preservação de evidências visa conservar, coletar e documentar com sucesso as evidências que podem contribuir para a compreensão do acidente. No entanto, mesmo que não possamos analisar totalmente as evidências (i.e., extrair todas as informações e explicações que gostaríamos) elas devem ser devidamente armazenadas para futura análise e referência. Além disso, o caso que se analisa hoje pode vir a ser o primeiro de uma série

que demonstra, por exemplo, a avaria de uma determinada IA que equipa um determinado modelo de automóvel autónomo ou os dados da XAU podem ser utilizados para testar hipóteses noutro caso.

No item 7.3 XAI, examinamos o emprego ex-ante da XAI. A seguir, focamos no emprego ex-post da XAI (como um conjunto de ferramentas forenses), i.e., como a XAI pode ser incorporada ao processamento de evidências digitais, nas etapas 6 (exame), 7 (análise) e 8 (apresentação), e como esses resultados, combinados com outros obtidos durante a Análise de Acidente, preenchem os requisitos legais para apuração da responsabilidade, fechando a lacuna entre nossas demandas legais e as explicações da XAI.

8.5 INTELIGÊNCIA ARTIFICIAL EXPLICATIVA FORENSE (IAEF)

Após percorrer essa árdua jornada, é de se esperar que a aplicação das técnicas de XAI se assemelhe à aplicação do teste Voigt-Kampff em Rachael por Deckard, i.e., um conjunto de perguntas para determinar sua natureza medindo funções como respiração, frequência cardíaca e dilatação pupilar em resposta a perguntas emocionalmente provocativas. Nossa tarefa é, de fato, aplicar uma série de testes, principalmente estatísticos, para explicar os dados e comportamentos da IA. Mas para a IAEF, nosso objetivo é responder como e por que as coisas aconteceram, reconstruir a história, ou seja, causalidade factual (o quê) e causalidade legal (como e por que), além de estabelecer fatores contribuintes e validar causas subjacentes.

Devemos ter uma definição clara do que estamos investigando; caso contrário, poderíamos nos apegar a um determinado cenário e procurar fatos que corroborem essa hipótese, apesar de evidências conflitantes. Além disso, do ponto de vista do queixoso, a explicação fornecida pela XAI pode ser paradoxal. As explicações ex-post costumam demonstrar a culpa do réu, mas também podem provar que outro infrator foi o responsável ou que a imprudência do querelante causou a perda ou dano, liberando o réu do dever de indenizar.

Por se tratar de uma especialização da Forense Digital, a metodologia da IAEF[75] segue as mesmas etapas descritas na Tabela 12 – Nove etapas do processamento de evidências. A seguir, focamos nas três etapas que requerem maior atenção: exame, análise e apresentação.

8.5.1 Exame

Se voltarmos aos nossos objetivos (especificamente reconstruir a história), é na etapa de Exame que escrevemos o roteiro. Aqui, é realizada uma busca sistemática aprofundada de evidências relacionadas ao possível crime, com foco na identificação e localização de evidências potenciais, possivelmente em locais não convencionais. Como resultado, construímos uma documentação detalhada para análise.

Na Ilustração 2 (um veículo que atingiu um caminhão que atravessou em sua frente), se perguntarmos o que aconteceu, a fim de mostrar a causalidade factual, podemos afirmar que o veículo com direção autônoma colidiu com um caminhão e causou o acidente. Para estabelecer a causalidade, devemos responder como e por que isso aconteceu.

Como isso aconteceu?

(1) O veículo autônomo **A** trafegava em uma via com um cruzamento.

(2) Um caminhão **C** atravessou em frente ao veículo **A**.

(3) O veículo autônomo **A** colidiu com o caminhão **C**.

Imagine por um momento que não temos nenhuma técnica XAI à nossa disposição – apenas técnicas forenses atuais. Com base nos vestígios da cena (e.g., marcas de frenagem, estilhaços etc.), em (1), das informações telemétricas contidas nos veículos – podemos supor que pelo menos o carro pilotado por IA tenha uma forma de módulo de controle de retenção – já extrairíamos evidências suficientes para estabelecer como o acidente ocorreu (e.g., aceleração durante o trajeto, comando para acionamento de freios etc.).

---

[75] Vide (PADOVAN, MARTINS e REED, 2022)

Em (2), quando reconstruímos o caminho percorrido por **C**, sabemos se ele fez a conversão de pista de forma correta, i.e., se ao entrar na frente de **A**, **C** causou o acidente.

Em (3), podemos estabelecer que **A** continuou seu caminho, e acabou colidindo com **C**. Mas sem XAI, não podemos estabelecer porque **A** não evitou a colisão, e.g., mudando sua rota.

Certamente, podemos conjeturar que **A** não 'percebeu' **C** uma vez que, do contrário, haveria marcas de pneu na pista. No entanto, ainda precisamos entender a raiz da decisão de **A**, ou podemos nos deparar com outros acidentes (possivelmente evitáveis) igualmente ou mais letais. Diversas perguntas emanam: "Por que **A** não percebeu **C**?", "Será que **A** teria tempo para evitar a colisão com **C**?", "Será que **A** só percebeu **C** tarde demais?"," Um motorista humano não teria percebido **C**?".

Estas e outras questões não podem ser respondidas sem as técnicas de XAI. Responder a essas perguntas é essencial para determinar as razões por trás de como o acidente ocorreu, tais como, quem errou e por que, e até mesmo para podermos comparar o comportamento da IA ao de um ser humano na situação.

Podemos simular o acidente com as informações obtidas na XAU e estabelecer o que o carro "percebeu" momentos antes do acidente, que comportamento ele previu para cada veículo, como (e com base em quê) decidiu continuar. Então, dependendo dos dados registrados e recuperados na XAU, imagens e vídeos do acidente podem ou não provar as afirmações de **C**.

Conforme exposto, temos duas tarefas distintas e igualmente importantes: estabelecer quais evidências teremos à nossa disposição para a fase de análise e elaborar o roteiro de avaliação que será aplicado às evidências.

Estabelecer quais evidências usaremos para a análise parece óbvio se tivermos os dados da XAU em mãos. No entanto, às vezes nos deparamos com situações em que recuperamos apenas parte (ou nenhum) desses dados. Outras vezes, o cruzamento de informações de diferentes evidências (não necessariamente digitais) sobre o mesmo evento é interessante ou mesmo necessário. Ocasionalmente, nos deparamos com situações em que precisamos aumentar nossos dados (mesmo

que tenhamos dados da XAU), compilar dados de casos semelhantes ou usar outros bancos de dados para executar determinados testes.

Por essa razão, é importante explicar a metodologia utilizada na fase de preparação dos dados, sempre atentos aos princípios descritos nas etapas 1 a 5 do processamento de evidências digitais (Tabela 12 – Nove etapas do processamento de evidências), com especial ênfase na justificação das fontes, documentando as modificações (linhagem de dados), e explicando ambos (ver item 7.3.1.2.1 Dados).

Para a tarefa de elaboração do roteiro de avaliação, precisamos estabelecer quais perguntas queremos responder, bem como delinear as áreas a serem abordadas e determinar quais informações devem ser extraídas das respostas.

Arrolar e tipificar as perguntas que queremos responder (i.e., positiva ou normativa) é o primeiro passo para a elaboração do roteiro. Lembre-se que nossas técnicas suportam ambos, mas as suposições são diferentes. Por exemplo, na Ilustração 2, poderíamos perguntar se o carro autônomo viu o caminhão envolvido (pergunta positiva) ou porque o carro autônomo não mudou para a faixa oposta (questão normativa).

Para responder à primeira pergunta, podemos extrair nossa explicação dos dados da XAU (i.e., com base nas amostras que encontramos) e responder sim ou não. Para a segunda pergunta, podemos extrair nossas respostas da IA com uma perspectiva ampla, (i.e., como o carro geralmente se comporta naquele tipo de situação) ou de uma perspectiva mais específica (i.e., porque o carro se comportou daquela maneira nesse caso específico). Nossas técnicas de explicação global, itens ❷e ❹ da Figura 13 – Principais fontes e tipos de explicação ex-post, atendem à demanda da pergunta ampla, enquanto a pergunta estrita exige o emprego de uma explicação local, itens ❸ e ❺ da Figura 13 – Principais fontes e tipos de explicação ex-post, avaliada usando dados da XAU aumentados ❻.

Para delinear as áreas a serem abordadas, agrupar as perguntas por finalidade se mostra interessante. Ao eliciar e dispor as perguntas, devemos sempre ter em mente seus objetivos, e.g., estabelecer o contexto da tarefa realizada pela IA e o ambiente em que foi realizada, ou entender por que – e não apenas o que – aconteceu em um determinado caso.

As informações obtidas, quer sobre os detalhes do acidente quer sobre algo não diretamente ligado, podem ser suficientes para responder a outras questões, levantar novas indagações, ou mesmo bastar para o que pretendemos e podemos analisar. Esse esclarecimento permite uma demanda de teste menor e mais eficiente e nos permite ver mais facilmente quais outros conjuntos probatórios precisamos para responder ao restante das nossas perguntas.

Mais tarde, correlacionamos as perguntas que queremos fazer com os testes que podemos fazer para respondê-las. Esse foco não apenas nos permite ser mais eficientes, mas também nos ajuda a estabelecer se precisamos de testes que expliquem a IA, avaliem o viés e a robustez ou se essas explicações exigirão outras técnicas.

As técnicas XAI a serem utilizadas variam de acordo com a tecnologia utilizada em cada um dos componentes auditados. Por exemplo, o módulo responsável pelo reconhecimento dos veículos e outros obstáculos, provavelmente se vale de uma das técnicas mais populares usadas para melhorar a precisão da classificação de imagens, a *Convolutional Neural Network* (CNN).

A CNN é uma rede neural que possui uma camada de convolução no início, em vez de alimentando a imagem inteira como uma matriz de números. A CNN divide a imagem em vários blocos (ou ladrilhos), reduzindo-os a uma forma mais fácil de ser processada. Esse processo de separação é importante quando estamos lidando com conjuntos de dados escaláveis ou massivos e crucial para evitar a perda de recursos e obter previsões precisas. A CNN então tenta prever a natureza de cada bloco e os alimenta em sua camada convolucional.

Ao jogar Imagem & Ação[76], porque o tempo é crítico, começamos nossos desenhos com recursos de baixo nível e posteriormente recursos de alto nível. Por exemplo, ao desenharmos uma ‘casa na praia’, começamos pelo bom e velho triangulo equilátero sobre um quadrado (para expressar uma casa) e depois partimos para as ondinhas e, se necessário o sol ou um sombreiro. Ou seja, formamos um conjunto de descrições que nos permite classificar (ou adivinhar) com confiança o que estamos “vendo”. Da mesma forma, quando nos pedem para descrever alguém

[76] Também conhecido como Pictionary fora do brasil.

que acabamos de conhecer, começamos com suas características principais, como formato do rosto, cor dos olhos e dos cabelos, e depois passamos para as idiossincrasias, como orelhas ligeiramente assimétricas ou o formato das sobrancelhas.

Da mesma forma, as CNN não estão limitadas a apenas uma camada convolucional. Formalmente, a primeira camada é responsável por capturar os recursos de baixo nível (e.g., bordas, cor, orientação e assim por diante) com as camadas subsequentes capturando recursos de alto nível. Depois de passar pelo processo acima, as camadas convolucionais produzem uma compreensão geral dos ladrilhos, semelhante à que os humanos fariam. Por fim, esses ladrilhos são alimentados em uma rede neural que tenta prever o que representa. Este processo permite que o computador paralelize as operações e detecte o objeto independentemente de onde ele possa estar localizado dentro da imagem.

Para mostrar os fatores contribuintes ou validar os fatores subjacentes, podemos usar a técnica de *Explainable Neural Network* (XNN) para explicar a CNN e entender quais recursos foram aprendidos, tornando-os explícitos pela visualização de recursos. Além disso, usando dissecação de rede, podemos vincular áreas altamente ativadas de canais da CNN com conceitos humanos (neste caso, humanos e primatas). Além disso, podemos usar técnicas de explicação de dados para medir distribuições em dados de treinamento, descobrindo assim uma possível fonte para o problema, bem como técnicas de imparcialidade para medir tendências e discrepâncias. Vale ressaltar que construir um roteiro de avaliação pode exigir várias iterações de preparação de dados e avaliação de perguntas. Na etapa seguinte, análise, é quando se consumarão todos os testes arrolados no roteiro.

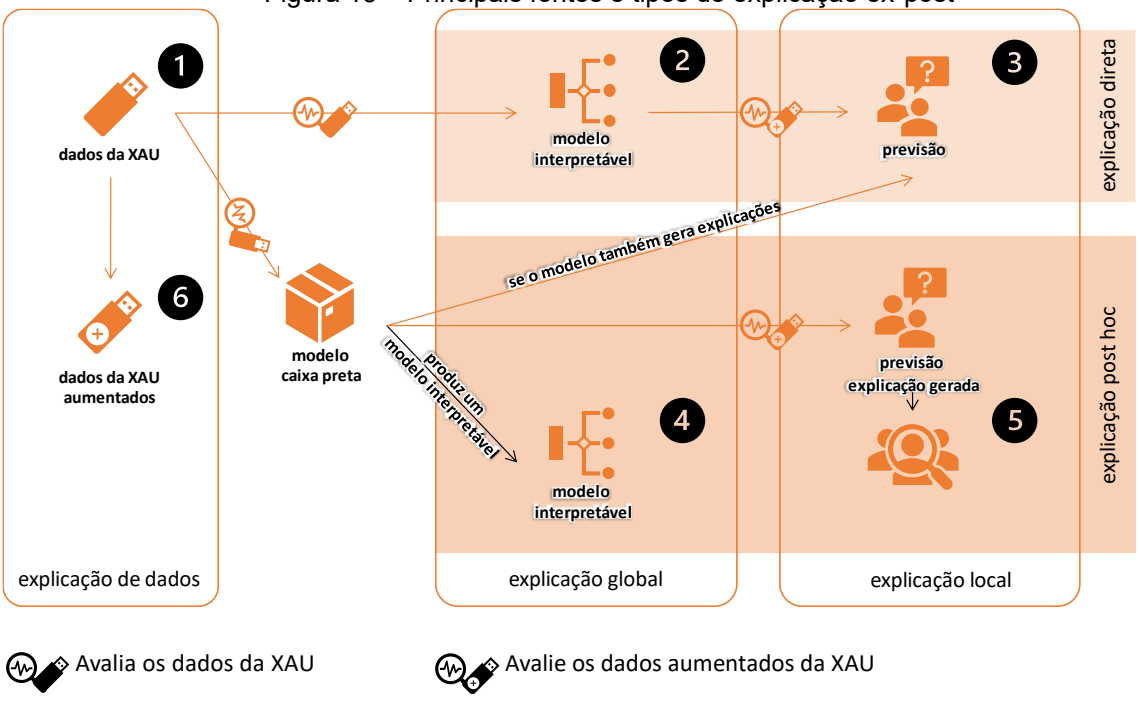
8.5.2 Análise

Na etapa anterior, nos preocupamos em reconstruir a história; aqui, estamos preocupados em responder aos porquês, estabelecer fatores contribuintes e validar os fatores subjacentes. Procuramos determinar a significância, explicar o modelo e os dados e tirar conclusões com base nas evidências encontradas.

Apoiar uma teoria do acidente pode levar várias iterações de investigação e análise (ensaios ou teste). A distinção neste caso é que esses ensaios podem não exigir altas habilidades técnicas para serem executados e, portanto, mais pessoas podem trabalhar no caso. Na verdade, muitos desses testes podem ser automatizados.

A Figura 13 – Principais fontes e tipos de explicação ex-post ilustra as principais fontes de explicação ex-post, i.e., onde e como proceder nossas análises para extrair as informações que respondem nossas perguntas. Podemos extrair nossas explicações dos dados da XAU ❶. Posteriormente, obtemos explicações sobre o modus operandi da IA ❷ e ❹ ou deliberações específicas ❸ e ❺. Como tal, as classificamos de acordo com sua origem (direta e post hoc) e abrangência (global e local).

Figura 13 – Principais fontes e tipos de explicação ex-post



Fonte: Padovan et al. (2022)

❶ Explicação dos dados da XAU – avalia as informações obtidas da XAU, mostrando o que aconteceu antes, durante e depois do acidente, i.e., uma espécie de lembrança testemunhal do acidente do ponto de vista da IA e o que ela previu em cada momento.

❷ Explicação Direta Global – avalia o processo de deliberação de um modelo explicável em uma visão holística, e.g., variáveis mais influentes e o comportamento primário.

❸ Explicação Direta Local – avalia uma predição individual (significativa) ou um conjunto de dados para testar hipóteses.

❹ Explicação Post hoc Global – avalia o processo de deliberação de uma caixa-preta, através de um modelo substituto, a partir de uma visão holística, e.g., variáveis mais influentes e o comportamento primário.

❺ Explicação Post hoc Local – avalia uma previsão individual ou um conjunto de dados para testar hipóteses e as explica.

❻ Dados Aumentados da XAU – uma cópia ampliada de uma amostra do conjunto de dados da XAU, i.e., com dados recém-criados ou cópias ligeiramente modificadas, para teste de hipótese.

Até aqui, discutido diversas técnicas, principalmente no item 7.3.1.2 Técnicas de Explicabilidade e em maiores detalhes no Anexo 1, que podem e são utilizadas nesta fase. Igualmente, abordamos no item 7.3.2.4 Limites limites e 7.3.2.5 Problemas problemas encontrados pelas técnicas de XAI. Por enquanto, nenhum algoritmo pode estimar se uma explicação provavelmente é uma racionalização ou capaz de detectar *fairwashing* por si só, e uma análise humana é essencial para detectar uma tentativa de fraude.

Este exemplo destaca a importância do processo holístico de análise e não apenas o uso indiscriminado de técnicas de exame de dados. A investigação deve analisar a tarefa que a IA estava realizando como um todo, i.e., ações e contexto, e.g., seu potencial para ter sucesso na tarefa, o processamento de informações, bar-

reiras projetadas (por exemplo, segurança cibernética e técnicas de IA de robustez), defesas administrativas, supervisão defesas e defesas culturais.

Agora, se a etapa de análise que acabamos de descrever pode ser retratada como uma montagem com várias cenas de experimentos científicos e uma estimulante trilha sonora, a próxima etapa (apresentação) é a notória cena do famoso detetive expondo o caso e revelando o culpado.

8.5.3 Apresentação

Na etapa anterior, nos preocupamos em determinar a significância, explicar o modelo e os dados e tirar conclusões com base nas evidências encontradas. Aqui, estamos preocupados com a *mise en scène*, ou seja, o arranjo de nossas conclusões na história reconstruída.

Aqui, resumimos e explicamos nossas descobertas em um relatório formal, permanente, auditável e defensável. Deve ser escrito em termos leigos usando terminologia abstrata (referenciando os detalhes específicos).

O profissional responsável pela apresentação deve ser capaz de determinar a melhor forma de desdobrar a informação. Por exemplo, para ilustrar os acidentes das ilustrações Ilustração 1 e Ilustração 2, um caminho simples, visual e acessível é desenvolver esboços e diagramas cronológicos identificando momentos-chave do acidente (locais, ações realizadas, lembranças de testemunhas etc.) destacando em cada um as deliberações da IA e suas consequências (sempre correlacionando com os dados coletados).

Como o foco da XAI é explicar a IA, o modelo e os dados, geralmente usamos técnicas de visualização de modelo e visualização de dados. Três aspectos fundamentais devem ser considerados ao desenvolver nossa apresentação de descobertas: o público-alvo, abstração de dados/tarefa e codificação.

Público-alvo – engloba um grupo de usuários-alvo, seu domínio de interesse, suas perguntas e seus dados. Por exemplo, um juiz precisa de respostas para

questões normativas que determinam a responsabilidade, enquanto um investigador de acidentes precisa de evidências e explicações.

Abstração de dados/tarefas – mapear as necessidades e dados do público-alvo em formas independentes do domínio. Por exemplo, na ilustração 2, um investigador de acidentes pode querer comparar o desempenho da IA em diferentes condições de iluminação, enquanto um juiz, comparar o desempenho da IA com um humano naquela condição específica.

Codificação - decidir sobre a maneira específica de abordar as tarefas listadas anteriormente. Por exemplo, o teste com diferentes condições de iluminação pode ser representado por um gráfico de barras ou uma fotomatriz com os diferentes resultados ilustrados, enquanto a comparação AI versus. humano, um único percentil ou uma frase “n vezes melhor” pode ser suficiente.

Ao final, todos esses resultados serão devidamente fundamentados e embalados em um laudo que fará parte do laudo pericial digital. Todas essas descobertas forenses serão incorporadas ao processo de ACR, que por sua vez gerará um relatório. O laudo da ACR será incorporado ao relatório do acidente, que por sua vez deverá apresentar, de forma fundamentada, o nexo causal de fato e de direito exigido pelo juízo.

O processo e as técnicas acima podem explicar os motivos que levaram a IA a tomar sua decisão, dado o relato de como ocorreu o acidente (através da XAU) e porque a perda ou dano então ocorreu (através de XAI).

9 CONCLUSÕES

Neste capítulo final, concluímos descrevendo os progressos realizados em direção à conformação do MAIA para dirimir questões de responsabilidade algorítmica e *accountability*. No item 9.1 Analisaremos os dois objetivos gerais, anteriormente definidos, e resumiremos as contribuições da pesquisa relatada. No tópico 9.2 abordaremos trabalhos futuros.

9.1 CONSIDERAÇÕES FINAIS E CONTRIBUIÇÕES DA PESQUISA

Esta tese teve como objetivos: (1) analisar, caracterizar e conformar *accountability* no âmbito da moral e da IA e (2) prover um cânone que auxilia as diversas partes interessadas (e.g., gestores, políticos, juristas, ativistas, cidadãos, consumidores etc.) a lidarem com os anseios por responsabilidade algorítmica (i.e., compreenderem, agirem, relatarem e – eventualmente – explicarem, consertarem e melhorarem os algoritmos que produzem, impactam ou que são afetados).

O arquétipo desenvolvido para exercer tais ofícios foi o Metamodelo de *accountability* para Inteligência Artificial (MAIA), que ocorre durante o ciclo de vida da IA (i.e., visa governar a conformação dos processos, produção, cadeia de suprimentos etc.), faculta as principais tarefas de *accountability* e é implementado pelo seguinte ciclo de decisão: Identificação das informações; Registro das informações; Análise das informações e Relato das informações (IRAR).

Para definir *accountability* no contexto da inteligência artificial e da moral, apresentamos, caracterizamos e diferenciamos conceitos de responsabilidade, responsabilidade civil, responsabilidade aquiliana e *accountability* na conjuntura sociopolítica normativa hodierna. E apresentamos as duas premissas relacionadas a *accountability*: exigibilidade (realizar certas ações, ou abster-se de) e responsividade (prestação de contas pelas ações realizadas).

Para a segunda meta, prover um cânone que auxilia os produtores a lidarem com os anseios por responsabilidade algorítmica dos stakeholders, desenvolvemos,

ao longo dessa produção, o Metamodelo de *accountability* para Inteligência Artificial (MAIA). O MAIA é baseado no metamodelo contábil; portanto, as nossas tarefas seguintes foram: expor o Metamodelo de *Accountability* (MA) e, ao longo da tese, apresentamos, adaptamos e comprovarmos sua aplicabilidade funcional, operacional e informacional para facultar as principais tarefas de *accountability* para IA.

O MA visa a prestação de contas de uma empresa, tanto na esfera pública como na privada, e consiste em quatro questões normativas implementadas por um ciclo decisório de quatro etapas. À vista disso, desenvolvemos cada uma delas.

Sempre que possível, deslindamos cada passo, apresentando uma abordagem interpessoal e interpartes, principais técnicas, fontes, justificativas e processo, além de uma equivalência entre a aplicação do MA na escrituração da empresa e na *accountability* da IA. Houve uma digressão quanto as técnicas de coleta e análise dos cômputos contábeis e informacionais; por isso, focamos apenas em apresentá-las para a IA quando abordamos XAI. No entanto, não houve perdas, dado que a dimensão finalista de cada ferramental se manteve.

Para alcançarmos nosso objetivo, precisávamos entender o quê estudávamos (i.e., a IA), não isoladamente, mas em uma estrutura socioambiental com características que evoluíram ao longo do tempo e estão sujeitas a uma série de causas e contingências (e.g., finalidade, lugar, cultura, história etc.). Vimos a necessidade de entender a dinâmica da IA em diversos contextos sociais, não apenas para exaurir qualquer dúvida sobre a real necessidade de *accountability*, mas para determinar princípios, normas e valores que nos são caros e devem ser incorporados à confecção da nossa solução.

Começamos definindo sete contextos diferentes: ética, moral, reputação, normas, leis, políticas e educação. Para cada um deles nós explicitamos os papéis, atividades, normas e valores. Por vezes, esses contextos se justapõem e pudemos perceber como a falta de entendimento por parte das pessoas sobre as ações da IA sempre foi ponto pacífico de atrito. Daí a necessidade de estudarmos no capítulo seguinte os stakeholders contemplados por nossos cômputos.

Logo após, estudamos a factibilidade de se inserir *accountability* no ciclo de vida da IA. Começamos pela segunda etapa do ciclo IRAR (Registro das informa-

ções). Para tal, apresentamos a *Explainable Accountable Unit* (XAU). Que consiste num sistema, ou um único módulo a depender da situação, que deve sempre registrar dados sobre a IA que posam ser usados para investigação de ocorrências, prestação de contas, depuração, imputabilidade e atende a requisitos, éticos, morais, reputacionais, normativos, legais, políticos e educacionais. Estabelecemos, também, quando (ex-ante ou ex-post) onde (diversas etapas do ciclo de vida da IA) e como (telemetria) coletar os dados que embasam nossos cômputos.

A seguir, averiguamos a terceira etapa do ciclo IRAR (Análise das informações), embasados pela QN3. Examinamos as principais técnicas de explicabilidade de IA. Foram dissertados dois grupos: explicabilidade, i.e., técnicas que visam aclarar a mecânica da IA, e justiça, i.e., técnicas que ambicionam dirimir vieses e injustiças nos algoritmos. Essas técnicas constituem um dos eixos que respalda o nosso objetivo de prover explicação.

Porém, de nada adiantava termos acesso a essas técnicas, se não pudéssemos acompanhar a IA durante sua vida útil. Vimos a necessidade de registrar os dados sobre a IA que nos possibilitassem regularmente explicar o que ela andava fazendo, como estava fazendo, por qual motivo estava fazendo e se estava sendo justa. Precisávamos de mecanismos e dados que pudessem ser usados para investigação de ocorrências, prestação de contas, depuração e imputabilidade.

Finalmente, averiguamos a quarta etapa do ciclo IRAR (Relato das informações), embasados pela QN4. Começamos por compreender como se dá o processo de apuração dos dados até se tornarem conhecimento. Ulteriormente, examinamos a técnica de Análise de Acidente e suas quatro etapas. Em seguida vimos as sete etapas da Análise da Causa Raiz e como ela, juntamente com a Forense Digital, promovem a preservação, coleta, validação, identificação, análise, interpretação, documentação e apresentação de evidências digitais, derivadas de fontes digitais, usadas para identificar as causas raiz de falhas ou defeitos.

9.2 TRABALHOS FUTUROS

Ao revermos os dois deveres principais da *accountability* (realizar certas ações, ou abster-se de, de acordo com as expectativas de um grupo e fornecer uma avaliação dessas ações para os interessados) a luz de tudo que discutimos até aqui, dimanam algumas inquirições sobre o atual, e futuro, emprego do MAIA para lidar com *accountability* algorítmica.

Pode-se perguntar se as explicações apresentadas são suficientes para determinar (e.g., em casos de perda ou dano resultado de uma decisão algorítmica) a responsabilidade da IA. A nosso ver, sim. Uma vez que os dados da XAU e as técnicas de XAI podem responder o quê, porque, quem, onde, quando e como, e, respondendo a essas questões, podemos estabelecer o nexo causal. Assim, atribuímos proporcionalmente responsabilidade a tais falhas e lidamos com problemas de *accountability* compartilhada e falta de conhecimento sobre os processos de decisão da IA.

A forma dos cômputos pode, e deve, ser adaptada para o público, mas a narrativa permanece a mesma. O léxico utilizado na descrição (i.e., as explicações) do processo apresentado a um juiz difere, por vezes substancialmente, do vocabulário utilizado para o júri ou adotado por técnicos. Os requisitos legais também podem nos obrigar a seguir formas predeterminadas de apresentação de informações, por exemplo, relatórios, transcrições, procedimentos e registros (e.g., relatórios de impacto da AAA22). O que é aceitável, ou exigido, depende do público e da finalidade e evoluirá com o tempo, o que é um tópico de acompanhamento interessante e necessário. Nos primeiros dias, exigiremos que seres humanos traduzam algumas explicações da XAI para o formato necessário. Ademais, novas ferramentas XAI podem ser concebidas para produzir diferentes explicações para públicos distintos sendo potencialmente um novo tópico de pesquisa.

Por falar em novas técnicas de XAI, atualmente dispomos de uma miríade, cada vez maior e mais especializada, de técnicas para explicar os mais variegados tipos de modelos e dados, e para diversos fins. São abrangentes as ênfases versadas (e.g., robustez, pegada ecológica, transparência etc.) e os temas abarcados

(e.g., veículos autônomos, energia, agricultura, aeronáutica etc.) por essas novas técnicas.

Já dispomos de *frameworks* de XAI que qualquer pessoa pode usar para interpretar um modelo de aprendizado de máquina, e.g., INNvestigate Neural Networks (ALBER, LAPUSCHKIN, *et al.*, 2019)[77], explAIner (SPINNER, SCHLEGEL, *et al.*, 2019)[78] e InterpetML (NORI, JENKINS, *et al.*, 2019)[79]. Essas e outras estruturas nos remetem a possibilidade, não muito remota, de *Explainable Artificial Intelligence as a Service* (XAIaaS) como um serviço a ser oferecido pelas diversas infraestruturas de *Artificial Intelligence as a Service* (AIaaS)[80].

Na verve de serviços inovadores, e possivelmente disruptivos, temos as diversas API oferecidas pela OpenAI (e.g., GPT-3[81], ChatGPT[82], DALL·E 2[83] etc.) embasando uma quantidade cada vez maior de aplicativos e sistemas hodiernos, variando de geradores de piadas a analista de logs; passando por geradores de peças teatrais e fornindo diversas ferramentas de geração e tratamento de imagem (OPENAI, 2023). Espera-se que essas API e suas versões futuras nos aproximem do mesmo nível de interação homem computador que se observa em séries de ficção científica como Star Trek (i.e., onde os tripulantes dirigem suas perguntas a "*computer*" e esta lhes responde o que foi indagado ou executa o que lhe foi pedido).

Essas e outras vindouras características e funcionalidades impactaram em novos pleitos, éticos, morais, reputacionais, normativos, legais, políticos e educacionais a serem impetrados pelos stakeholders. No tocante a novas demandas, num futuro próximo alguns dos projetos de lei que tratamos (e.g., AAA22) tornar-se-ão leis. Como discorrido, a legislação trará uma série de novas demandas, e provavel-

---

77 Um pacote Python expansível fácil de usar que implementa uma grande variedade de métodos de explicação visual.
78 Uma estrutura unificada que ajuda os usuários a entenderem os modelos de aprendizado de máquina usando diferentes técnicas explicáveis.
79 Uma biblioteca Python de código aberto com muitos algoritmos de interpretabilidade, que podem ser facilmente integrados ao código.
80 Refere-se a ferramentas e serviços que permitem às empresas implementarem e dimensionarem IA por uma fração do custo de uma solução interna.
81 Generative Pre-trained Transformer 3, é um modelo de processamento de linguagem natural que requer uma pequena quantidade de texto de entrada para gerar grande, relevante e sofisticado volume de texto.
82 Um modelo que interage de forma conversacional, i.e., responde a perguntas subsequentes, admite seus erros, questiona premissas incorretas e rejeita solicitações inapropriadas.
83 Novo sistema de IA que pode criar imagens e arte realistas a partir de uma descrição em linguagem natural.

mente tornarão uma série de pleitos arrazoados na QN1, obrigatórios, impactando não só na obrigatoriedade e conformação dos cômputos, bem como, na realização de certas ações (e.g., um responsável humano pela efetivação dos resultados da IA) e renúncia a outras (e.g., proibição de emprego da IA para executar certas tarefas) por parte dos produtores e responsáveis pela IA.

Será que no futuro ainda falaremos de *algorithmic accountability*? Se sim, Como o MAIA se comportará em cenários como estes? A priori, basta reiterar as quatro questões normativas e reimplantá-las (i.e., adaptando ou redigindo um novo modelo). Como vimos, para implementar seu modelo de processo de negócio você precisa descrever as atividades do processo (MA), as estruturas de dados (ME) e as regras de negócios (RN) que restringem e orientam as operações do processo.

Como as Regras de Negócios impactam diretamente nos Modelos de Atividade e Modelos Estruturais usitados, ao reiteramos as 4 questões normativas, elicitamos novas restrições e orientações para novos processos. As temáticas abordadas irão evoluir com as demandas futuras. Provavelmente explicabilidade e imparcialidade serão pontos consuetudinários, não por perderem relevância e sim por termos aprendido a “lidar” com eles de forma satisfatória, cabendo a posteridade lidar com reinvindicações futuras.

Às futuras gerações, o lembrete: *algorithmic accountability* é, em última instância, um pacto firmado entre pessoas para proceder e explicar, segundo aspectos éticos, morais, reputacionais, normativos, legais, políticos, educacionais etc., algoritmos que produzem, impactam ou são afetados.

9.3 LIMITAÇÕES

Cabe ainda discutirmos algumas das limitações do MAIA quanto a sua validade e aplicabilidade. Discutiremos como a primeira está intimamente ligada a segunda.

Pela falta de uma especificação formal, não apresentamos nessa obra nenhuma prova formal da corretude ou completude do metamodelo. Como mencionado

em 1.5 Metodologia, o MAIA herda sua validação finalista (i.e., identificar, registrar, analisar e relatar informações) do metamodelo contábil. No entanto, sua validação ampla caberia ao longo de um continuum entre uma validação explícita (e.g., um sistema, uma prova formal etc.), e uma validação tácita (e.g., coletânea de modelos derivados do MAIA que atendem as diversas demandas informacionais da responsabilidade algorítmica).

Quanto a aplicabilidade, o MAIA sempre estará restrito à nossa capacidade de prover respostas às suas 4 questões normativas ou implementar suas principais tarefas. Em 1.5 Metodologia, explicitamos o nosso entendimento sobre a abrangência do termo *algorithmic accountability* para além dos algoritmos de inteligência artificial que envolvem aprendizado de máquina e isso por si só já nos traz uma quase dicotomia entre o que podemos ou não analisar.

Por exemplo, na década de 1970, os pesquisadores de IA perceberam que para que seus sistemas resolvessem problemas reais de forma satisfatória, era necessário incorporar grandes quantidades de conhecimento sobre o problema. Isso levou ao surgimento da “Engenharia do Conhecimento”, que busca formas de viabilizar a utilização de conhecimentos de especialistas na solução de problemas complexos. A tecnologia resultante desse campo de estudos é chamada Sistemas Especialistas e ainda é amplamente usado em aplicações comerciais (Russell & Norvig, 2010).

Na contemporaneidade, o processo de prestação de contas e prover explicabilidade sobre um determinado sistema especialista é trivial, quando uma determinada metodologia para desenvolvimento de software é seguida. Certamente, qualquer que seja a metodologia usitada, ela empregará alguma forma de identificação e registro dos dados e informações tratados pelo sistema. As regras embutidas no sistema derivam de especialista e foram devidamente registradas e validadas em algum documento, trivializando os processos de análise e relato na maioria dos casos.

No entanto, se pegarmos o Chat GPT[84] (GPT significa *Generative Pre-trained Transformer* ou Transformador Generativo Pré-treinado em português), podemos

---

84 Um tipo de IA treinada em grandes quantidades de dados usando aprendizado auto supervisionado que emprega aprendizado profundo para produzir respostas semelhantes à seres humanos.

nos deparar com desafios homéricos ao tentarmos prestar contas ou prover explicabilidade sobre uma determinada resposta sua. Por exemplo, no tocante a identificar e registrar as informações, o ChatGPT foi temporariamente banido na Itália, no início de 2023, devido a preocupações com privacidade, uma vez que coletava ilegalmente dados pessoais de usuários e não possuía um sistema de verificação de idade para impedir que menores fossem expostos a material ilícito (SATARIANO, 2023) e seus produtores estão sendo processados por violação de direitos autorais (METZ, 2022).

No quesito explicabilidade, sim, existem várias técnicas de XAI que podem ser usitadas com GPTs. Algumas das técnicas mais populares incluem “Visualização de atenção”, que é um método para visualizar os pesos de atenção do modelo (MACHLEV, HEISTRENE, *et al.*, 2022), “Propagação de relevância em camadas” (LRP), que é um método para atribuir importância aos recursos de entrada de uma modelo (KOHLBRENNER, BAUER, *et al.*, 2020) e DeepLift, recursos importantes de aprendizado profundo, (MACHLEV, HEISTRENE, *et al.*, 2022). Outras técnicas de XAI que podem ser empregadas, abordadas no Anexo 1, incluem LIME, Valores Shapley, Contrafactuais e demais técnicas agnósticas. Mas e quando nos deparamos com indagações sobre manifestações genuínas de inteligência[85] e reais riscos para a humanidade[86]? Para algumas dessas, não possuímos sequer uma definição para abordar o problema, quanto mais uma forma de relatar a resposta.

O que podemos extrair desses breves exemplos explanados até aqui é que sem uma devida identificação e registro das informações empregadas pela IA, quer seja para sua confecção, depuração, operação etc., não podemos garantir rastreabilidade e responder a questões referentes a dados (*quem*, *o quê*, *onde* e *quando* e que variam em temática de vieses a copyright passando por privacidade) e, eventualmente, evoluir nossa análise. Além disso, sempre estaremos limitados a aquilo que podemos compreender e explicar.

---

85 Vide (FERREIRA, 2023), (ALMENARA, 2023).
86 Vide (BBC NEWS BRASIL, 2023).

REFERÊNCIAS

2001: A Space Odyssey. Direção: Stanley Kubrick. [S.l.]: [s.n.]. 1968.

ACKOFF, R. L. From data to wisdom. **Journal of Applied Systems Analysis 16**, p. 3–9, 1989.

ACQUAVIVA, M. C. **Dicionário Acadêmico de Direito**. 2ª ed. ed. São Paulo: Editora Jurídica Brasileira, 2001.

AFP. Boris Johnson deseja novo acordo de divórcio da UE. **Estado de Minas**, 25 July 2019. Disponivel em: <https://www.em.com.br/app/noticia/internacional/2019/07/25/interna_internacional,1072344/boris-johnson-deseja-novo-acordo-de-divorcio-da-ue.shtml>.

AIESEC. O que é e quais as vantagens do planejamento estratégico empresarial? **AIESEC**, 12 December 2021. Disponivel em: <https://aiesec.org.br/o-que-e-e-quais-as-vantagens-do-planejamento-estrategico-empresarial/#:~:text=O%20planejamento%20estrat%C3%A9gico%20empresarial%20%C3%A9,planos%20de%20execu%C3%A7%C3%A3o%20e%20metas.>.

AÏVODJI, U. et al. **Fairwashing:** the risk of rationalization. 36 th International Conference on Machine. Long Beach: PMLR. 2019.

ALBER, M. et al. iNNvestigate neural networks! **J. Mach. Learn. Res.**, 20, n. 93, 2019. 1–8.

ALBRECHT, W. S. et al. **Accounting:** Concepts & Applications, 10e. 10th. ed. Mason: Thomson South-Western, 2008. 1339 p. ISBN 978-0-324-37615-4.

ALMENARA, I. Como o ChatGPT mentiu para convencer um humano a trabalhar para ele. **Canaltech**, 31 March 2023. Disponivel em: <https://canaltech.com.br/inteligencia-artificial/como-o-chatgpt-mentiu-para-convencer-um-humano-a-trabalhar-para-ele-245170/>.

ALTABBAKH, H. et al. STAMP – Holistic system safety approach or just another risk model? **Journal of Loss Prevention in the Process Industries**, p. 109-119, 2014.

AMAZON. Amazon Super Bowl Commercial 2020 - #BeforeAlexa. **YouTube**, 2 February 2020. Disponivel em: <https://www.youtube.com/watch?v=RF9t2rFmTVE>.

AMENO, F.; DIAS, T. ALEXA É 'SOLUÇÃO DE ESCUTA ATIVA', DIZ EMPRESÁRIO QUE VENDE SOFTWARES DE VIGILÂNCIA A POLÍCIAS E FORÇAS ARMADAS. **The intercept Brasil**, 18 April 2022. Disponivel em: <https://theintercept.com/2022/04/18/alexa-e-solucao-de-escuta-ativa-diz-empresario-que-vende-softwares-de-vigilancia-a-policias-e-forcas-armadas/>.

AMNESTY INTERNATIONAL. Protecting the right to equality and non-discrimination in machine learning systems. **The Toronto Declaration**, 16 May 2018. Disponivel em: <https://www.torontodeclaration.org/declaration-text/english/>.

ANDREWS, D. A.; KIESSLING, R. D.; MICKUS, S. G. The risk principle of case classification: An outcome evaluation with young adult probationers. **Canadian Journal of Criminology**, v. 4, n. 28, p. 377-384, 1986.

ANDREWS, D.; BONTA, J. **LSI-R: The Level of Service Inventory-Revised.** Toronto. 1995.

ANDREWS, D.; BONTA, J. **The psychology of criminal conduct**. Cincinnati: Anderson, 2003.

ANGELES, P. A. **Harper Collins Dictionary of Philosopy**. New York: Harper Collins, 1992.

ARBULU, R. Amazon está ouvindo conversas entre usuários e Alexa, diz agência. **Canaltech**, 11 April 2019. Disponivel em: <https://canaltech.com.br/espionagem/amazon-esta-ouvindo-conversas-entre-usuarios-e-alexa-diz-agencia-136939/>.

AUSTRALIAN GOVERNMENT. Privacy Act 1988. **Federal Register of Legislation**, 13 Dezembro 2019. Disponivel em: <https://www.legislation.gov.au/Details/C2020C00025>.

BAIRD, C.; HINES, C.; BEMUS, B. **The Wisconsin Case Classification/Staff Deployment Project: Two-Year Follow-Up Report**. Madison, WI. 1979.

BARROS, M. Ex-funcionário do Google afirma que redes sociais tratam usuários como produtos. **Olhar Digital**, 4 October 2021. Disponivel em: <https://olhardigital.com.br/2021/10/04/internet-e-redes-sociais/ex-funcionario-do-google-afirma-que-redes-sociais-tratam-usuarios-como-produtos/>.

BARTON, A. D. Public sector accountability and commercial-in-confidence outsourcing contracts. **Accounting, Auditing & Accountability Journal, 19(2)**, p. 256–71, 2006.

BBC NEWS. Quem é Boris Johnson, o polêmico novo premiê britânico que conduzirá Brexit. **BBC News**, 23 July 2019. Disponivel em: <https://www.bbc.com/portuguese/internacional-49066508>.

BBC NEWS. Twitter investigates racial bias in image previews. **BBC News**, 21 Setembro 2020. Disponivel em: <https://www.bbc.com/news/technology-54234822>.

BBC NEWS BRASIL. Inteligência artificial: o alerta de mil especialistas sobre 'risco para a humanidade'. **BBC News Brasil**, 30 March 2023. Disponivel em: <https://www.bbc.com/portuguese/articles/c89yywnx5lyo>.

BELSLEY, D. A.; KUH, E.; WELSH, R. E. **Regression Diagnostics:** Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons, 1980.

BERK, R. et al. Fairness in Criminal Justice Risk Assessments: The State of the Art. **Sociological Methods & Research**, 2017.

BERNARDINO, A. L. Ex-engenheira da Google preocupada com "armas autónomas" capazes de "atrocidades em massa". **MAGG**, 15 September 2019. Disponivel em: <https://magg.sapo.pt/atualidade/atualidade-internacional/artigos/ex-engenheira-da-google-preocupada-com-armas-autonomas-capazes-de-atrocidades-em-massa>.

BETSY MORRIS. Parents' Dilemma: When to Give Children Smartphones. **The Wall Street Journal**, 12 January 2018. Disponivel em: <https://www.wsj.com/articles/iphones-vs-parents-the-tug-of-war-over-americas-children-1515772695?mod=article_inline>.

BLACK Mirror; Season 3 Episode 1 - Nosedive. Direção: Joe Wright. [S.l.]: [s.n.]. 2016.

BLADE Runner. Direção: Ridley Scott. [S.l.]: [s.n.]. 1982.

BLOCH, H. P. Successful failure analysis strategies. **Reliability Advantage:** Training Bulletin, 7 May 2005. Disponivel em: <http://www.heinzbloch.com/docs/ReliabilityAdvantage/Reliability_Advantage_Volume_3.pdf>.

BLOCH, H. P. Structured failure analysis strategies solve pump problems. **Machinery Lubrication**, 21 August 2011. Disponivel em: <http://www.machinerylubrication.com/Read/28467/pump-failure-analysis>.

BLODGET, H. Sex Without A Condom Is "Rape" In Sweden. **Business Insider**, 8 December 2010. Disponivel em: <https://www.businessinsider.com/latest-on-julian-assanges-sex-crimes-sex-without-a-condom-is-rape-in-sweden-says-a-swede-2010-12>.

BLOOMBERG. Facebook Suspends Trump Election Data Firm for Policy Breach. **TIME**, 17 March 2018. Disponivel em: <https://time.com/5204387/facebook-suspends-cambridge-analytica>.

BONTA, J. L. Native inmates: Institutional response, risk, and needs. **Canadian Journal of Criminology**, p. 49-62, 1989.

BONTA, J.; MOTIUK, L. L. Classification to halfway houses: A quasi-experimental evaluation. **Criminology**, n. 28, p. 497-506, 1990.

BPMN. Object Management Group Business Process Model and Notation. **BPMN 2.0**, 2017. Disponivel em: <http://www.omg.org/spec/BPMN/2.0/>.

BRADESCO. BIA. A Inteligência Artificial do Bradesco. **YouTube**, 11 August 2018. Disponivel em: <https://www.youtube.com/watch?v=k3brZzuC5Ug>.

BRASIL. Lei Nº 13.709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais (LGPD)**, Brasília,DF, 14 August 2018. Disponivel em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>. Acesso em: 26 February 2020.

BRASIL. Lei Nº 13.834, de 4 de junho de 2019. **Lei da fake news**, 4 June 2019. Disponivel em: <http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Lei/L13834.htm>. Acesso em: 26 February 2020.

BRASIL. Projeto de Lei n° 5051, de 2019. **Estabelece os princípios para o uso da Inteligência Artificial no Brasil.**, 2019. Disponivel em: <https://www25.senado.leg.br/web/atividade/materias/-/materia/138790>. Acesso em: 26 February 2020.

BRASIL. ACESSIBILIDADE NA CÂMARA - Brecha na lei. **Câmara dos Deputados**, 15 January 2022. Disponivel em: <https://www2.camara.leg.br/a-camara/estruturaadm/gestao-na-camara-dos-deputados/responsabilidade-social-e-ambiental/acessibilidade/glossarios/dicionario-de-libras/b/brecha-na-lei#:~:text=Lacuna%20ou%20oportunidade%20na%20lei.>.

BURT, R. S.; KNEZ, M. Kinds of Third-Party Effects on Trust. **Rationality & Society**, p. 255-92, 1995.

BUSINESSDICTIONARY.COM. Accountability. **Business Dictionary**, 2 November 2017. Disponivel em: <www.businessdictionary.com/definition/accountability.html>.

CALABRICH, B. Conceito(s) de norma Uma breve análise sobre a classificação de von Wright. **Revista de Informação Legislativa**, Junho 2008. 55-62.

CALMON, F. P. et al. **Optimized Pre-Processing for Discrimination**. 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: [s.n.]. 2017.

CÂMARA, I. Lawtech: o que é e como está o mercado para essas startups? **Startse**, 20 March 2018. Disponivel em: <https://www.startse.com/noticia/startups/lawtech/o-que-e-lawtech>.

CAMPBELL, A.; GLASS, K. C. The Legal Status of Clinical and Ethics Policies, Codes, and Guidelines in Medical Practice and Research. **McGill law journal**, p. 473-89, 2001.

CAPLAN, R. et al. Algorithmic Accountability: A Primer. **Data & Society**, 18 Abril 2018. Disponivel em: <https://datasociety.net/wp-content/uploads/2019/09/DandS_Algorithmic_Accountability.pdf>.

CARDOSO, C. Tesla melhora desempenho das portas do Modelo X… com um update via software. **MeioBit**, 12 Outubro 2016. Disponivel em: <https://www1.tecnoblog.net/meiobit/2016/tesla-os-8-melhora-tempo-de-acionamento-das-portas-do-model-x-wings/>.

CARRIER, B. **Defining Digital Forensic Examination and Analysis Tools.** 2002 Digital Forensics Research Workshop. [S.l.]: [s.n.]. 2002.

CASALS, A. New Technologies in Surgery. In: OLLERO, A., et al. **ROBOT 2017:** Third Iberian Robotics Conference, Volume 2. Geneve: Springer, 2017. p. 537-541.

CERUZZI, P. E. **A History of Modern Computing 2nd**. Cambridge: The MIT Press, 2003.

CGTI/ACS. TJDFT usa inteligência artificial para aprimorar sistemas. **TJDFT**, May 2019. Disponivel em: <https://www.tjdft.jus.br/institucional/imprensa/noticias/2019/maio/tjdft-usa-inteligencia-artificial-para-aprimorar-sistemas>.

CHAPMAN, P. et al. [S.l.]: SPSS, 2000.

CHAPPELL, B. 'It Was Installed For This Purpose,' VW's U.S. CEO Tells Congress About Defeat Device. **NPR**, 8 October 2015. Disponivel em: <https://www.npr.org/sections/thetwo-way/2015/10/08/446861855/volkswagen-u-s-ceo-faces-questions-on-capitol-hill>.

CHOLLET, F. Building Autoencoders in Keras. **The Keras Blog**, 14 May 2016. Disponivel em: <https://blog.keras.io/building-autoencoders-in-keras.html>.

CHOULDECHOVA, A. **Fair Prediction with Disparate Impact:** A Study of Bias in Recidivism Prediction Instruments. Big Data (2016). [S.l.]: [s.n.]. 2016.

CHOWDHURY, M.; APON, A.; DEY, K. **Data Analytics for Intelligent Transportation Systems**. Lieden: Elsevier, 2017.

CLOUSER MCCANN, P. ; SHIPAN, C. R. How many major US laws delegate to federal agencies? (almost) all of them. **Political Science Research and Methods**, v. 10, n. 2, p. 438–444, 2022.

CNN. Polanski arrested in connection with 1970s sex charge. **CNN**, 28 September 2009. Disponivel em: <http://edition.cnn.com/2009/CRIME/09/27/zurich.roman.polanski.arrested/>.

COHEN, W. W. Fast effective rule induction. **Machine Learning Proceedings**, p. 115-123, 1995.

COMISSÃO EUROPÉIA. COM(2021) 206 final 2021/0106(COD). **Proposta de regulamento do parlamento europeu e do conselho que estabelece regras harmonizadas em matéria de inteligência artificial (regulamento inteligência artificial) e altera determinados atos legislativos da união**, 21 April 2021. Disponivel em: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>.

CONSELHO FEDERAL DE MEDICINA. **Código de Ética Médica:** Resolução CFM nº 2.217, de 27 de setembro de 2018, modificada pelas Resoluções nº 2.222/2018 e 2.226/2019. Brasília: Gráfica Marina Ltda, 2019.

COOK, R.. Detection of Influential Observation in Linear Regression. **Technometrics**, p. 15-18, 1977.

COOLING, E.; HERBERS, P. V. Considerations in Autopilot Litigation. **Journal of Air Law and Commerce**, n. 48, p. 693-923, 1983.

CORBETT-DAVIES, S. et al. **Algorithmic Decision Making and the Cost of Fairness**. Proceedings KDD'17. [S.l.]: [s.n.]. 2017.

COURSERA. Coursera. **Coursera**, 02 February 2020. Disponivel em: <https://www.coursera.org/search?query=AI&>.

D'ALESSANDRO, B.; O'NEIL, C.; LAGATTA, T. Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. **Big Data**, p. 120-134, 2017.

DICKSON, B. Inside DARPA's effort to create explainable artificial intelligence. **TechTalks**, 10 January 2019. Disponivel em: <https://bdtechtalks.com/2019/01/10/darpa-xai-explainable-artificial-intelligence/>.

DINARDI, G. 15 Steps to Create a Killer Communication Plan. **Nextiva**, 26 September 2019. Disponivel em: <https://www.nextiva.com/blog/steps-create-communications-plan-template.html>.

DN/LUSA. Cambridge Analytica teve papel decisivo no referendo do Brexit. **Diário de Notícias**, 27 March 2018. Disponivel em: <https://www.dn.pt/mundo/brexit-cambridge-analytica-teve-papel-decisivo-no-referendo-diz-ex-funcionario-9217257.html>.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. **arXiv**, 8 March 2017. Disponivel em: <https://arxiv.org/abs/1702.08608v2>.

DUNBAR, R. **How Many Friends Does One Person Need?** New York: Harvard University Press, 2010.

DW. Parlamento britânico aprova lei que impede Brexit sem acordo. **Deutsch Welle**, 4 September 2019. Disponivel em: <Parlamento britânico aprova lei que impede Brexit sem acordo>.

DWORK, C. et al. **Fairness Through Awareness**. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. [S.l.]: [s.n.]. 2012.

ELIOT, T. S. **The Rock**. London: Faber & Faber, 1934.

EMANUEL, E. J. et al. Attitudes and Practices of Euthanasia and Physician-Assisted Suicide in the United States, Canada, and Europe. **JAMA**, p. 316(1):79–90, 2016.

ENDSLEY, M. R. Toward a Theory of Situation Awareness in Dynamic Systems. **Human Factors: The Journal of the Human Factors and Ergonomics Society**, v. 37, n. 1, p. 32–64, 1995.

ESTADOS UNIDOS DA AMÉRICA. The Age Discrimination in Employment Act. **U.S. Equal Employment Opportunity Commission (EEOC)**, 15 December 1967.

Disponivel em: <https://www.eeoc.gov/statutes/age-discrimination-employment-act-1967>.

ESTADOS UNIDOS DA AMÉRICA. Fair Housing Act. **The United States of America Department of Justice**, 11 April 1968. Disponivel em: <https://www.justice.gov/crt/fair-housing-act-1>.

ESTADOS UNIDOS DA AMÉRICA. Equal Credit Opportunity Act. **USC Code House**, 23 March 1976. Disponivel em: <http://uscode.house.gov/view.xhtml?req=granuleid%3AUSC-prelim-title15-chapter41-subchapter4&edition=prelim>.

EUROPEAN ORGANIZATION FOR CIVIL AVIATION EQUIPMENT. **EUROCAE ED 112**. Saint-Denis. 2013.

EUROPEAN PARLIAMENT. European Parliament. **artificial intelligence:** questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice, 20 January 2021. Disponivel em: <https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_EN.html>.

EUROPEAN PARLIAMENT AND OF THE COUNCIL. Regulation (EU) 2016/679. **General Data Protection Regulation (GDPR)**, 27 April 2016. Disponivel em: <https://gdpr-info.eu/>. Acesso em: 26 February 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, Providence, 17, n. 3, 27 July 1996. 37-54.

FEDERAL AVIATION ADMINISTRATION. 14 CFR 23.1329. **Aeronautics and Space**, 1 January 2002. Disponivel em: <https://www.govinfo.gov/app/details/CFR-2002-title14-vol1/CFR-2002-title14-vol1-sec23-1329>. Acesso em: 26 February 2020.

FEDERAL AVIATION ADMINISTRATION. 14 CFR 91.3. **Responsibility and authority of the pilot in command**, 1 January 2006. Disponivel em:

<https://www.govinfo.gov/app/details/CFR-2006-title14-vol2/CFR-2006-title14-vol2-sec91-3>. Acesso em: 26 February 2020.

FEDERAL AVIATION ADMINISTRATION. **TSO-C124b**. Washington, DC. 2019.

FERNANDES, D. A. Logística legal:“A lei que não pega”. Por quê?. **Profanações**, v. 3, n. 1, p. 5-19, 2016.

FERREIRA, Y. Inteligência artificial finge ser humana, contrata pessoa e espanta cientistas. **Forum**, 31 March 2023. Disponivel em: <https://revistaforum.com.br/ciencia-e-tecnologia/2023/3/31/inteligncia-artificial-finge-ser-humana-contrata-pessoa-espanta-cientistas-133651.html>.

FISCHBACH, N. Biometrics in 2019: Increased Security or New Attack Vector? **Threat Post**, 9 Janeiro 2019. Disponivel em: <https://threatpost.com/biometrics-in-2019-increased-security-or-new-attack-vector/140683/>.

FISHER, D. H. The Ethics of Artificial Intelligence. **The Ethics of Artificial Intelligence**, 31 January 2020. Disponivel em: <https://my.vanderbilt.edu/aiethics/>.

FLORIDI, L. et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. **Minds and Machines**, p. 689–707, 2018.

FOOT, P. The Problem of Abortion and the Doctrine of the Double Effect. **Oxford Review**, 1967.

FREEMAN, E. R.; REED, D. L. Stockholders and Stakeholders: A New Perspective on Corporate Governance. **California Management Review**, p. 88-106, 1983.

FREEMAN, R. E. **Strategic management:** A stakeholder approach. Boston: Pitman, 1984.

FRIEDLER, S. A.; SCHEIDEGGER, C. E.; VENKATASUBRAMANIAN, S. On the (im)possibility of fairness. **ArXiv**, 23 September 2016. Disponivel em: <https://arxiv.org/abs/1609.07236>.

G1. Brexit: União Europeia concorda em adiar saída do Reino Unido para 31 de janeiro de 2020. **G1**, 28 October 2019. Disponivel em:

<https://g1.globo.com/mundo/noticia/2019/10/28/brexit-uniao-europeia-concorda-em-adiar-saida-do-reino-unido-para-31-de-janeiro-de-2020.ghtml>.

GALHOTRA, S.; BRUN, Y.; MELIOU, A. **Fairness Testing:** Testing Software for Discrimination. Proceedings of ESEC/FSE'17. [S.l.]: [s.n.]. 2017.

GARITSELOV, O.; MOHANTY, S.; KOUGIANOS, E. A Comparative Study of Metamodels for Fast and Accurate Simulation of Nano-CMOS Circuits. **IEEE Transactions on Semiconductor Manufacturing (TSM)**, 25, n. 1, Fevereiro 2012. 26–36.

GEBER RAMALHO. Oficina de Ética e Inteligência Artificial. **Oficina de Ética e Inteligência Artificial**, December 2019. Disponivel em: <https://cin.ufpe.br/~etica/>.

GERMANN, A. C.; DAY, F. D.; GALLATI, R. R. J. **Introduction to Law Enforcement and Criminal Justice**. [S.l.]: [s.n.], 1985.

GOETZ, A. M.; JENKINS, R. Hybrid forms of accountability: citize engagement in institutions of public sector oversight in India. **Public Management Review**, 3, n. 3, September 2001.

GOGGIN, C.; GENDREAU, P.; GRAY, G. Associates and social interaction. **Forum on Corrections Research**, v. 3, n. 10, p. 24-27, 1998.

GOH, BRENDA. China to bar people with bad 'social credit' from planes, trains. **Reuters**, 16 March 2018. Disponivel em: <https://www.reuters.com/article/us-china-credit/china-to-bar-people-with-bad-social-credit-from-planes-trains-idUSKCN1GS10S>.

GOLDSTEIN, A. et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. **Journal of Computational and Graphical Statistics**, v. 24, n. 1, p. 44-65, 2015.

GOOGLE. Sobre. **Google**, 02 Janeiro 2022. Disponivel em: <https://about.google/intl/pt-BR/>.

GRANVILLE, K. Como a Cambridge Analytica recolheu dados do Facebook. **Folha de São Paulo**, 21 March 2018. Disponivel em:

<https://www1.folha.uol.com.br/mercado/2018/03/como-a-cambridge-analytica-recolheu-dados-do-facebook.shtml>.

GRAY, R.; ADAMS, C.; OWEN, D. **Accountability, Social Responsibility and Sustainability**. Harlow: UK, 2014.

GREENEMEIER, L. 20 Years after Deep Blue: How AI Has Advanced Since Conquering Chess. **Scientific American**, 20 Dezembro 2021. Disponivel em: <https://web.archive.org/web/20211220224453/https://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/>.

GRIFFIN, M.. **Contabilidade e finanças**. São Paulo: Saraiva, 2012.

GROVE, W. M. et al. Clinical versus mechanical prediction: A meta-analysis. **Psychological Assessment**, v. 1, n. 12, p. 19–30, 2000.

GUIMÓN, P.; SAHUQUILLO, M. R. 'Brexit' vence e Reino Unido deixará a União Europeia. **El País**, 24 June 2016. Disponivel em: <https://brasil.elpais.com/brasil/2016/06/24/internacional/1466741749_403437.html>.

GUINNESS WORLD RECORDS LIMITED. Robert Wadlow: Tallest man ever. **Guinness World Records**, 27 Agosto 1955. Disponivel em: <https://www.guinnessworldrecords.com.br/records/hall-of-fame/robert-wadlow-tallest-man-ever#:~:text=Os%20g%C3%AAmeos%20ent%C3%A3o%20nomearam%20Robert,ft%2011%2C1%20pol).>.

GUINNESS WORLD RECORDS LIMITED. Heaviest man ever. **Guinness World Records**, 4 Abril 1978. Disponivel em: <https://www.guinnessworldrecords.com.br/world-records/heaviest-man>.

GUNNING, D. **DARPA's explainable artifcial intelligence (XAI) Program**. Proceedings of the 24th International Conference on Intelligent User Interfaces. Sanibel Island: IUI. 2019. p. 47.

GUNNING, D.; AHA, D. W. DARPA's Explainable Artifcial Intelligence (XAI) Program. **AI Magazine**, v. 40, n. 2, p. 44-58, 2019.

GURUMOORTHY, K. S. et al. **Efficient Data Representation by Selecting Prototypes with Importance Weights**. IEEE International Conference on Data Mining (ICDM). Beijing: IEEE. 2019. p. 260-269.

HARARI, Y. N. **21 Lessons for the 21st Century**. New York: Random House, 2008.

HARDT, M.; PRICE, E.; SREBRO, N. Equality of Opportunity in Supervised Learning. **Advances in Neural Information Processing Systems**, 2016.

HARDT, M.; PRICE, E.; SREBRO, N. **Equality of Opportunity in Supervised Learning**. 30th Conference on Neural Information Processing Systems (NIPS 2016). Barcelona: [s.n.]. 2016.

HARPER, G.; PICKETT, S. Methods for mining HTS data. **Drug Discovery Today**, v. 11, n. 694, p. 15–16, 2006.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning - Data Mining,Inference,and Prediction**. New York: Springer, 2017.

HAUGELAND, JOHN, J. **Artificial Intelligence:** The Very Idea. Cambridge: MIT Press, 1985.

HAWKINS, A. J. Uber is at fault for fatal self-driving crash, but it's not alone. **The Verge**, https://www.theverge.com/2019/11/19/20972584/uber-fault-self-driving-crash-ntsb-probable-cause, 19 November 2019. Disponivel em: <https://www.theverge.com/2019/11/19/20972584/uber-fault-self-driving-crash-ntsb-probable-cause>.

HERN, A. Cambridge Analytica did work for Leave.EU, emails confirm. **The Guardian**, 30 July 2019. Disponivel em: <https://www.theguardian.com/uk-news/2019/jul/30/cambridge-analytica-did-work-for-leave-eu-emails-confirm>.

HERN, ALEX. Cambridge Analytica did work for Leave.EU, emails confirm. **The Guardian**, 30 July 2019. Disponivel em: <https://www.theguardian.com/uk-news/2019/jul/30/cambridge-analytica-did-work-for-leave-eu-emails-confirm>.

HIGH-LEVEL EXPERT GROUP ON AI. Ethics guidelines for trustworthy AI. **Digital Single Market**, 8 April 2019. Disponivel em: <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419>.

HINTON, G. ; SALAKHUTDINOV, R. R. Reducing the Dimensionality of Data with Neural Networks. **Science**, 313, 28 July 2006. 504-507.

HOLLNAGEL, E. **FRAM:** The Functional Resonance Analysis Method: Modelling Complex Socio-technical Systems. [S.l.]: CRC Press, 2012.

HOLTE, R. C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. **Machine Learning**, p. 63-91, 1993.

HORTON , HELENA. Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours. **The Telegraph**, 24 March 2016. Disponivel em: <https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/]>.

HSU, F.; CAMPBELL, M.; HOANE, A. J. **Deep Blue System Overview**. ICS '95: Proceedings of the 9th international conference on Supercomputing. [S.l.]: [s.n.]. 1995. p. 240–244.

IBM. What is explainable AI? **Explainable AI (XAI)**, 12 March 2021. Disponivel em: <https://www.ibm.com/watson/explainable-ai>.

IEEE. **IEEE nº 610.12**. New york. 1990.

IEEE. **IEEE Std 1044**. New York. 2009.

IGNACIO, L. Fisco diferencia multa previdenciária para órgãos públicos e empresas. **Valor Econômico**, 05 July 2012. Disponivel em: <https://valor.globo.com/legislacao/noticia/2012/07/05/fisco-diferencia-multa-previdenciaria-para-orgaos-publicos-e-empresas.ghtml>.

INTERNATIONAL CIVIL AVIATION ORGANIZATION. **Operation of Aircraft**. Montreal. 2010.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 14040:2006**. Genebra. 2006.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/IEC JTC 1/SC 42 Artificial intelligence**. Genebra. 2020.

ISO. **ISO/IEC 25000**. Genève. 2005.

ISO. **ISO/IEC 24765**. Genève. 2009.

ISRANI, E.. When an Algorithm Helps Send You to Prison. **The New York Times**, 26 October 2017. Disponivel em: <https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html>.

ITU. **Key ICT indicators for developed and developing countries and the world (totals and penetration rates)**. New York. 2015.

JASHAPARA, A. **Knowledge Management:** an Integrated Approach. Harlow: FT Prentice Hall, 2005.

JOBIN, A.; IENCA, M.; VAYENA, E. The global landscape of AI ethics guidelines. **Nature Machine Intelligence**, p. 389-399, 2019.

JOHNSON, G.; SCHOLES, K. **Exploring Corporate Strategy**. 3rd. ed. Englewood Cliffs: Prentice-Hall, 1993.

JÜTTING, J. **Institutions and Development: A Critical Review**. Paris. 2003.

KAMIRAN, F.; CALDERS, T. Data preprocessing techniques for classification. **Knowl Inf Syst**, p. 1-33, 2012.

KAMISHIMA, T. et al. **Fairness-Aware Classifier with Prejudice Remover Regularizer**. European Conference, ECML PKDD 2012. Bristol: Springer. 2012. p. 35-50.

KAUFMAN, D. Consumo de energia e emissão de CO2 dos algoritmos de inteligência artificial: como evitar uma catástrofe climática. **Época Negócios**, 18 December 2020. Disponivel em: <https://epocanegocios.globo.com/colunas/IAgora/noticia/2020/12/consumo-de-energia-e-emissao-de-co2-dos-algoritmos-de-inteligencia-artificial-como-evitar-uma-catastrofe-climatica.html>.

KETTINGER, W.; GROVER, V. **Process Think:** Winning Perspectives For Business Change in the Information Age. [S.l.]: IDEA Group Publishing Inc, 2000.

KILBERTUS, N. et al. **Avoiding Discrimination Through Causal Reasoning**. In Advances in Neural Information Processing Systems. [S.l.]: [s.n.]. 2017.

KIM, B.; KHANNA, R.; KOYEJO, O. **Examples are not Enough, Learn to Criticize!** 29th Conference on Neural Information Processing Systems (NIPS). Barcelona: NIPS. 2016. p. 1-11.

KING v Great Britain China Centre, 11 October 1991.

KIRKPATRICK, D. D. Perfis russos tentaram influenciar votação do 'brexit', indicam estudos. **Folha de São Paulo**, 16 Nobember 2017. Disponivel em: <https://www1.folha.uol.com.br/mundo/2017/11/1935898-perfis-russos-tentaram-influenciar-votacao-do-brexit-indicam-estudos.shtml>.

KLEINBERG, J. M.; MULLAINATHAN, S.; RAGHAVAN, M. **Inherent Trade-Offs in the Fair Determination of Risk Scores**. ITCS. [S.l.]: [s.n.]. 2017.

KLOEPFFER, W. Life Cycle Sustainability Assessment of Products. **Int J Life Cycle Asses**, v. 13, n. 2, p. 89 – 95, 2008.

KOHLBRENNER, M. et al. Towards Best Practice in Explaining Neural Network Decisions with LRP. **arxiv.org**, 13 July 2020. Disponivel em: <https://arxiv.org/abs/1910.09840>.

KOLSTAD, C. D.; ULEN, T. S.; JOHNSO, G. V. **The Theory and Practice of Command and Control in Environmental Policy**. Abingdon: Routledge, 2003. 331–344 p.

KOPP, B. **Telemetry Systems Engineering**. 1. ed. Massachusetts: Artech House, 2002. 493-524 p. ISBN 1580532578.

KUMAR, A.; SATTIGERI, P.; AVINASH, B. **Variational Inference of Disentagled Latent Concepts From unlabeled Observations**. ICLR 2018. Vancouver: ICLR. 2018.

KUMAR, B. **An illustrated dictionary of aviation**. New York: McGraw-Hill, 2005.

KUSNER, M. J. et al. **Counterfactual Fairness**. In Advances in Neural Information Processing Systems. [S.l.]: [s.n.]. 2017.

LANGE, D.; LEE, P.; DAI, Y. Organizational Reputation: A Review. **Journal of Management**, p. 153-184, 2011.

LAURA AND JOHN ARNOLD FOUNDATION. Public Safety Assessment (PSA). **Public Safety Assessment (PSA)**, 07 February 2020. Disponivel em: <https://www.psapretrial.org/about>.

LAURA AND JOHN ARNOLD FOUNDATION. Public Safety Assessment (PSA). **Risk factors and formula**, 15 January 2020. Disponivel em: <https://www.psapretrial.org/about/factors>.

LEE, PETER. Learning from Tay's introduction. **Official Microsoft Blog**, 25 March 2016. Disponivel em: <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.00000gjdpwwcfcus11t6oo6dw79gw>.

LETHAM, B. et al. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. **The Annals of Applied Statistics**, p. 1350-1371, 2015.

LIA, T. Websites crippled as copyright war gets personal. **The Sydney Morning Harold**, 21 January 2012. Disponivel em: <https://www.smh.com.au/technology/websites-crippled-as-copyright-war-gets-personal-20120120-1qa8k.html>.

MACHIAVELLI, N. **Discorsi sopra la prima deca di Tito Livio.** Firenze: [s.n.], 1531.

MACHLEV, R. et al. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. **Energy and AI**, 9, August 2022. 100-169.

MADHANI, P.. Role of Voluntary Disclosure and Transparency in Financial Reporting. **CORPORATE FINANCIAL REPORTING - CHANGING SCENARIO**, p. 75-81, 2008.

MCCLELLAND, D. **The Achieving Society**. New Jersey: Princeton, 1961.

MCTIC. **EBIA - Estratégia Brasileira de Inteligência Artificial**. Brasília. 2021.

METZ, C. Lawsuit Takes Aim at the Way A.I. Is Built. **The New York Times**, 23 November 2022. Disponivel em: <https://www.nytimes.com/2022/11/23/technology/copilot-microsoft-ai-lawsuit.html>.

MICHELLE MA. The Future of Everything - How to teach kids about ai. **The Wall Street Journal**, 13 May 2019. Disponivel em: <https://www.wsj.com/articles/how-to-teach-kids-about-ai-11557759541>.

MITCHELL, R. K.; AGLE, B. R.; WOOD, D. J. Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of. **The Academy of Management Review**, p. 853-886, 1997.

MOJSILOVIC, A. Introducing AI Explainability 360. **Introducing AI Explainability 360**, 4 June 2021. Disponivel em: <https://www.ibm.com/blogs/research/2019/08/ ai-explainability-360/>.

MORRIS, C. Brexit: A três semanas do prazo, ainda é possível adiar a saída do Reino Unido da União Europeia? **BBC News**, 10 March 2019. Disponivel em: <https://www.bbc.com/portuguese/internacional-47422962>.

MURRAY-WEBSTER, R.; SIMON, P. Making Sense of Stakeholder Mapping. **Published in PM World Today**, VIII, n. 11, 1 November 2006. 1.

MUSCHARA, T. INPO's approach to human performance in the United States commercial nuclear power industry. **IEEE Xplore Digital Library**, 2007. Disponivel em: <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4413179&url=http%3A%2F%2Fieeexplore.ieee.or>.

NABI, R.; SHPITSER, I. **Fair Inference On Outcomes**. AAAI. [S.l.]: [s.n.]. 2018.

NASA. **Telemetry: Summary of concept and rationale**. [S.l.], p. 89. 1987.

NATIONAL TRANSPORTATION SAFETY BOARD OFFICE OF HIGHWAY SAFETY. **Crash Summary HWY18FH011**. Mountain View, CA. 2019.

NATIONAL TRANSPORTATION SAFETY BOARD OFFICE OF HIGHWAY SAFETY. **Automation And Data Summary Factual Report HWY18FH011**. Mountain View, CA. 2020.

NEW YORK TIMES. NEW DEVICE MAKES AIRSHIPS FOOLPROOF. **The New York Times**, 27 November 1916. 20. Disponivel em: <https://www.nytimes.com/1916/11/27/archives/new-device-makes-airships-foolproof-capt-jv-martin-applies-for.html>.

NEWCOMBE, R. From client to project stakeholders: a stakeholder mapping approach. **Construction Maangement and Economics**, 21 December 2003. 841-848.

NIILER, E. Can AI Be a Fair Judge in Court? Estonia Thinks So. **WIRED**, 25 March 2019. Disponivel em: <https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/>.

NORI, H. et al. InterpretML: A Unified Framework for Machine Learning Interpretability. **arxiv.org**, 19 Setembro 2019. Disponivel em: <https://doi.org/10.48550/arXiv.1909.09223>.

NOWAK, M.; SIGMUND, K. Evolution of indirect reciprocity. **Nature**, n. 437, p. 1291–1298, 2005.

NTBS. **Accident Report NTSB/AAR-10/01 PB2010-910401**. New York. 2009.

O GLOBO. Facebook: Zuckerberg pode ser apontado como principal culpado por vazamentos de dados de usuários. **O Globo**, 19 April 2019. Disponivel em: <https://oglobo.globo.com/economia/facebook-zuckerberg-pode-ser-apontado-como-principal-culpado-por-vazamentos-de-dados-de-usuarios-23611291>.

OECD. OECD. Council Recommendation on Artificial Intelligence. **OECD.AI Policy Observatory**, 2019 May 2019. Disponivel em: <https://www.google.com/url?client=internal-element-cse&cx=014590198432315495341:fw0os7p8avk&q=https://oecd.ai/assets/files/OECD-LEGAL-0449-

en.pdf&sa=U&ved=2ahUKEwjO0rOb9aH8AhWrgpUCHfoCBBkQFnoECAIQAQ&usg=AOvVaw1_JNaJF0su4ErYXGjfp_Ie>.

OECD. Recommendation of the Council on Artificial Intelligence. **OECD Legal Instruments**, 5 May 2019. Disponivel em: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

OPENAI. GPT3 Demo. **GPT3**, 2 Janeiro 2023. Disponivel em: <https://gpt3demo.com/product/gpt-3>.

OSOWSKY, J. et al. **Projeção 2D em conjuntos de imagens médicas usando a teoria de range image.** III CLAEB. João Pessoa: IFMBE. 2004. p. 1307-1310.

O'SULLIVAN, A.; SHEFFRIN, S. M. **Economics:** Principles in Action. New Jersey: Pearson Prentice Hall, 2003.

OSWALD, V. 'Como na guerra, não há vencedores no Brexit', diz especialista. **O Globo**, 14 April 2019. Disponivel em: <https://oglobo.globo.com/mundo/como-na-guerra-nao-ha-vencedores-no-brexit-diz-especialista-23596209>.

PADOVAN, P. H.; MARTINS, C. M.; REED, C. Black is the new orange: how to determine AI liability. **Artifcial Intelligence and Law**, 2022.

PALMER, G. **A Road Map for Digital Forensic Research. Technical Report DTR-T001-01, DFRWS, Report From the First Digital Forensic Research Workshop (DFRWS)**. [S.l.]. 2001.

PASQUALE, F. **The Black Box Society:** The Secret Algorithms That Control Money and Information. Boston: Harvard University Press, 2016.

PAYNE, B. H. An Ethics of Artificial Intelligence Curriculum for Middle School Students. **An Ethics of Artificial Intelligence Curriculum for Middle School Students**, 5 August 2019. Disponivel em: <https://docs.google.com/document/d/1e9wx9oBg7CR0s5O7YnYHVmX7H7pnITfoDxNdrSGkp60/edit#heading=h.ictx1ljsx0z4>.

PDPC. MODEL ARTIFICIAL INTELLIGENCE GOVERNANCE FRAMEWORK. **Personal Data Protection Commission**, 21 January 2020. Disponivel em: <http://go.gov.sg/ai-gov-mf-2>.

PEDRESCHI, D.; RUGGIERI, S.; TURINI, F. **A study of top-k measures for discrimination discovery**. Proceedings of the 27th Annual ACM Symposium on Applied Computing. Trento: Association for Computing Machinery. 2012. p. 126–131.

PIATETSKY-SHAPIRO, G. Methodology Poll. (2007). **KDnuggets**, 14 June 2020. Disponivel em: <https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm>.

PICKOVER, C. A. **Artificial Intelligence:** An Illustrated History : from Medieval Robots to Neural Networks. Estados Unidos: Sterling, 2019.

PORTO EDITORA. bode expiatório na Infopédia. **Infopédia**, 08 jan. 2022. Disponivel em: <https://www.infopedia.pt/apoio/artigos/$bode-expiatorio>.

PROPUBLICA. How We Analyzed the COMPAS Recidivism Algorithm. **ProPublica**, 23 May 2016. Disponivel em: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

QUINSEY, V. L. et al. **Violent offenders:** Appraising and managing risk. Washington, DC: American Psychological Association, 1998.

RASMUSSEN, J. Risk management in a dynamic society: a modelling problem. **Safety Science**, p. 183-213, 1997.

REINO UNIDO. Sex Discrimination Act 1975. **Legislation.gov.uk**, 12 November 1975. Disponivel em: <https://www.legislation.gov.uk/ukpga/1975/65/enacted>.

REINSEL, D.; GANTZ, J.; RYDNING, J. **The Digitalization of the World - From Edge to Core**. Framingham, p. 28. 2018.

REITH, M.; CARR, C.; GUNSCH, G. An Examination of Digital Forensic Models. **International Journal of Digital Evidence**, v. 1, n. 3, 2002.

RENO, R. ; CIALDINI, R. B.; KALLGREN, C. A. The transsituational influence of social norms. **Journal of Personality and Social Psychology**, 1993. 104-112.

REZENDE, F. C. Por que reformas administrativas falham? **Revista Brasileira de Ciências Sociais**, 50, Outubro 2002.

RFI. Boris Johnson é confirmado premiê do Reino Unido e garante Brexit até 31 de outubro. **RFI**, 23 July 2019. Disponivel em: <http://www.rfi.fr/br/europa/20190723-boris-johnson-e-confirmado-premie-do-reino-unido-e-garante-brexit-ate-31-de-outubro>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"Why should I trust you?:** Explaining the predictions of any classifier.". 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco: ACM. 2016.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **Anchors:** High-Precision Model-Agnostic Explanations. AAAI Conference on Artificial Intelligence (AAAI). New Orleans: AAAI. 2018.

RICH, D. L. **Queen Bess:** Daredevil aviator. Washington & London: Smithsonian Institution Press, 1993.

RIEDESEL, J. **Software Telemetry - Reliable Loggin and Monitoring**. Shelter Island: Manning Publications Co., 2021.

RIEFFEL, E. G.; POLAK, W. H. **Quantum Computing:** A Gentle Introduction. Reino Unido: MIT Press, 2014.

RODRIGUES, O. C.; MU, G. B. "Modo cão": atualização da Tesla deixa o carro mais seguro para os cachorros. **Auto Esporte**, 11 Março 2019. Disponivel em: <https://autoesporte.globo.com/carros/noticia/2019/03/modo-cao-atualizacao-da-tesla-deixa-o-carro-mais-seguro-para-os-cachorros.ghtml>.

ROLLAND, C. **A Comprehensive View of Process Engineering**. Proceedings of the 10th International Conference on Advanced Information Systems Engineering. London: Springer-Verlag. 1998. p. 1-24.

RTP. Brexit. Parlamento britânico aprova acordo de saída. **RTP Notícias**, 20 December 2019. Disponivel em: <https://www.rtp.pt/noticias/mundo/brexit-parlamento-britanico-aprova-acordo-de-saida_n1193378>.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence:** A Modern Approach. 3ª. ed. Hoboken: Pearson, 2010. ISBN 978-0-13-604259-4.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence:** A Modern Approach. 4ª. ed. Hoboken: Pearson, 2021. ISBN 9780134610993.

SAMPLE, I. Computer says no: why making AIs fair, accountable and transparent is crucial. **The Guardian**, 5 November 2017. Disponivel em: <https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial>.

SAS INSTITUTE. Amsterdam UMC uses the SAS® Platform and AI solutions to increase speed and accuracy of tumor evaluations. **SAS**, 1 Maio 2021. Disponivel em: <https://www.sas.com/pt_br/customers/amsterdam-umc.html>.

SATARIANO, A. ChatGPT Is Banned in Italy Over Privacy Concerns. **The New York Times**, 31 March 2023. Disponivel em: <https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html#:~:text=Italy%E2%80%99s%20data%20protection%20authority%20said%20OpenAI%2C%20the%20California,ban%20ChatGPT%20as%20a%20result%20of%20privacy%20concerns.>.

SAVULESCU, J.; MASLEN, H. Moral Enhancement and Artificial Intelligence: Moral AI? In: ROMPORTL, J.; ZACKOVA, E.; KELEMEN, J. **Beyond Artificial Intelligence, Topics in Intelligent Engineering and Informatics**. New York: Springer International Publishing, 2015. p. 79-95.

SCHECK, W. Lawrence Sperry: Genius on Autopilot. **Aviation History Magazine**, November 2004. 27. Disponivel em: <https://www.historynet.com/lawrence-sperry-autopilot-inventor-and-aviation-innovator.htm>.

SCHEDLER, A. **The Self-Restraining State:** Power and Accountability in New Democracies. London: Lynne Rienner Publishers, 1999. 13–28 p. ISBN 978-1-55587-773-6.

SCHIEK, D.; WADDINGTON, L.; BELL, M. **Cases, Materials and Text on National, Supranational and International Non-Discrimination Law**. [S.l.]: Hart Publishing, 2007.

SCHMANDT-BESSERAT, D. **Before Writing**. Texas: University of Texas Press, v. I, 1992.

SCHNEIER, B. **Liars and Outliers**. Indianapolis: John Wiley & Sons, 2012.

SENADO FEDERAL. Suplemento (nº B) - DSF nº 204. **DSF**, 8 December 2022. Disponivel em: <https://legis.senado.leg.br/diarios/BuscaDiario?tipDiario=1&datDiario=09/12/2022&paginaDireta=3&indSuplemento=Sim&codSuplemento=B&desVolumeSuplemento=&desTomoSuplemento=>.

SILVER, D. et al. Mastering the game of Go with deep neural networks and tree search. **Nature**, (7587), 28 Janeiro 2016. 484–489.

SIMONITE, T. When It Comes to Gorillas, Google Photos Remains Blind. **Wired**, 1 nov. 2018. Disponivel em: <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.

SKEEM, J. L.; LOWENKAMP, C. Risk, Race, & Recidivism: Predictive Bias and Disparate Impact. **SSRN**, 14 June 2016. Disponivel em: <https://ssrn.com/abstract=2687339>.

SPINNER, T. et al. explainer: A visual analytics framework for interactive and explainable machine learning. **IEEE transactions on visualization and computer graphics**, 26, n. 1, 2019. 1064–1074.

STURGEON, L.; HARMAN, G.; THOMSON, J. J. Thomson Against Moral Explanations. **Philosophy and Phenomenological Research**, March 1998. 199-206.

SULLINS, ; DIGNUM , V.; CHOWDHURY, R. Responsible Innovation in the Age of AI. **Responsible Innovation in the Age of AI**, 2020. Disponivel em: <https://ieeexplore.ieee.org/servlet/opac?mdnumber=EW1496>.

SULLIVAN, M. THE SWING RIOTS. **University of Kent**, November 2017. Disponivel em: <https://www.thinkautomation.com/histories/automation-and-the-swing-riots/>.

SWETS, J. A.; DAWES, R. M.; MONAHAN, J. Psychological science can improve diagnostic decisions. **Psychological Science in the Public Interest**, p. 1–26, 2000.

THE NEW YORK TIMES. 'Double Irish With a Dutch Sandwich'. **The New York Times**, 28 April 2012. Disponivel em: <https://archive.nytimes.com/www.nytimes.com/interactive/2012/04/28/business/Double-Irish-With-A-Dutch-Sandwich.html>.

THE Terminator. Direção: James Cameron. [S.l.]: [s.n.]. 1984.

TIMES. No Hands. **Times**, p. 63, 1947.

TOEWS, R. AI Will Transform The Field Of Law. **Forbes**, 19 December 2019. Disponivel em: <https://www.forbes.com/sites/robtoews/2019/12/19/ai-will-transform-the-field-of-law/#583e82917f01>.

TRAUTMAN, P. S. A Computer Pioneer Rediscovered, 50 Years On. **The New York Times**, 20 abr. 1994. Disponivel em: <https://web.archive.org/web/20161104051054/http://www.nytimes.com/1994/04/20/news/20iht-zuse.html>.

UDEMY. Udemy. **Udemy**, 7 February 2020. Disponivel em: <https://www.udemy.com/courses/search/?src=ukw&q=AI>.

UML. UML 2.0 Superstructure. **UML Website**, 2005. Disponivel em: <https://www.uml.org/>.

UNITED STATES. H.R.2231 - Algorithmic Accountability Act of 2019. **CONGRESS.GOV**, 10 April 2019. Disponivel em: <https://www.congress.gov/bill/116th-congress/house-bill/2231/text>. Acesso em: 26 February 2020.

UNITED STATES. H.R.6580 - Algorithmic Accountability Act of 2022. **CONGRESS.GOV**, 3 February 2022. Disponivel em: <https://www.congress.gov/bill/117th-congress/house-bill/6580/text>.

VAN DER MAATEN, L.; HINTON, G. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, p. 2579-2605, 2008.

VASILAKOS, A. V.; CHANG, W. **Molecular Computing:** Towards a Novel Computing Architecture for Complex Problem Solving. Alemanha: Springer International Publishing, 2014.

VERMA, S.; RUBIN, J. **Fairness Definitions Explained**. 2018 ACM/IEEE International Workshop on Software Fairness. Gothenburg: IEEE. 2018. p. 1-7.

VITASEK, K. Supply Chain Management Terms and Glossary. **CSCMP**, 2019 April 2019. Disponivel em: <https://cscmp.org/sites/default/files/user_uploads/resources/downloads/glossary-2013.pdf?utm_source=cscmpsite&utm_medium=clicklinks&utm_content=glossary&utm_campaign=GlossaryPDF>.

VOLD, L. Strict Liability for Aircraft Crashes and Forced Landings on Ground Victims Outside of Established Landing Areas. **Hastings Law Journal**, p. 1-33, 1953.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. **Harvard Journal of Law & Technology**, p. 841-893, 2018.

WALKER, K. An external advisory council to help advance the responsible development of AI. **Google Blog**, 26 March 2019. Disponivel em: <https://www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/>.

WILSON, P. F.; DELL, L. D.; ANDERSON, G.. **Root Cause Analysis:** A Tool for Total Quality Management. Milwaukee: ASQ Quality Press, 1993. ISBN 0-87389-163-5.

WILSON, P. J. Filcher of Good Names: An Enquiry Into Anthropology and Gossip. **Man, New Series**, p. 93-102, 1914.

WOOLLASTON, V. How do Google's driverless cars work? **Alphr**, 04 April 2016. Disponivel em: <https://www.alphr.com/cars/7038/how-do-googles-driverless-cars-work>.

WRIGHT, K.; CLEAR, T.; DICKSON, P. Universal application of probation risk assessment instruments: A critique. **Criminology**, v. 1, n. 22, p. 113-134, 1984.

WRIGHT, K.; CLEAR, T.; DICKSON, P. Universal application of probation risk assessment instruments: A critique. **Criminology**, v. I, n. 22, p. 113-134, 1984.

WZAMEN01. HP computers are racist. **YouTube**, 10 December 2009. Disponivel em: <https://www.youtube.com/watch?v=t4DT3tQqgRM>.

YEOM, S.; TSCHANTZ, M. C. Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews, 2018.

YONG, E. A Popular Algorithm Is No Better at Predicting Crimes Than Random People. **The Atlantic**, 17 January 2018. Disponivel em: <https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/>.

ZEMEL, R. et al. **Learning Fair Representations**. Proceedings of the 30th International Conference on Machine Learning, PMLR. Atlanta: JMLR. 2013. p. 325-33.

ZHANG, B. H.; LEMOINE, B.; MITCHELL, M. Mitigating Unwanted Biases with Adversarial Learning. **AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society**, 22 January 2018. 335–340.

ŽLIOBAITė, I. A survey on measuring indirect discrimination in machine learning. **ArXiv**, 31 Octorber 2015. Disponivel em: <https://arxiv.org/abs/1511.00148>.

倪, 雪莹. **多地将拒服兵役者**纳入征信“**黑名**单”. **新京报**, 19 March 2018. Disponivel em: <http://www.bjnews.com.cn/news/2018/03/19/479533.html>.

記者林宜樟. **國道插爆貨車車廂！特斯拉變穿雲箭畫面曝光**. **Liberty Times News**, 1 Junho 2020. Disponivel em: <https://news.ltn.com.tw/news/society/breakingnews/3183411>.

ANEXO A – PRINCIPAIS TÉCNICAS DE XAI

O seguinte anexo visa expor as principais técnicas de explicabilidade (XAI) e imparcialidade (viés).

As ferramentas de XAI serão classificadas quanto a origem, i.e., dados ou modelo (direta ou indireta). Resultado, especificidade e escopo serão abordados dentro de cada método. Também é fornecido uma expectativa do nível de compreensão que a técnica prove.

Uma categoria especial (dentro de modelo) intituladas “exemplos”, foi criada para compreender métodos que – apesar de poderem ser incluídos nas outras categorias – deveriam figurar juntos, devido a sua finalidade.

As ferramentas de mitigação de vieses serão classificadas quanto a população, fase de aplicação, visão e metas. Também são fornecidos uma descrição e expectativa das metas que a técnica almeja.

A.1 EXPLICAÇÃO DE DADOS

Os modelos de aprendizagem de máquina têm sua gênese nos dados usados em seu treinamento. Assim, é benéfico compreender as características e atributos desses dados antes de começar o treinamento. Para tal, veremos técnicas de visualização, compreensão e higienização desses dados.

As principais técnicas de projeção 2D são: Non-Linear Principal Components Analysis (n-LPCA), Multidimensional Scaling (MDS), t-Distributed Stochastic Neighbor Embedding (t-SNE), e Autoencoder Networks.

A.1.1 Non-Linear Principal Components Analysis (n-LPCA)

O algoritmo Non-Linear Principal Components Analysis (n-LPCA) é uma generalização não linear do PCA que usa uma rede "codificadora" adaptável e multinível para transformar os dados de alta dimensão em um código de baixa dimensão e uma rede "decodificadora" semelhante para recuperar os dados do código (HINTON e SALAKHUTDINOV, 2006).

Especificidade: Genérico - Pode ser usada para visualizar conjuntos de dados complexos com muitas variáveis.

Escopo: Local e global. Provê uma visão geral ou visualizações granulares de partes locais do conjunto de dados.

Compreensão: Média - É usado para confirmar os resultados modelados pela IA; para confirmar se a IA está classificando corretamente - i.e., quando se conhece as hierarquias, classes ou cluster dos dados de treinamento ou teste; para confirmar se atributos e estruturas similares são projetados relativamente perto e atributos dispares longe uns dos outros.

A.1.2 Multidimensional Scaling (MDS)

O Multidimensional Scaling (MDS) é um meio de visualizar o nível de similaridade de casos individuais em um conjunto de dados. O MDS é usado para representar, i.e., n pontos mapeados em um espaço cartesiano, informações sobre as distâncias entre um conjunto de n objetos (HASTIE, TIBSHIRANI e FRIEDMAN, 2017).

Especificidade: Genérico - Pode ser usada para visualizar conjuntos de dados complexos com muitas variáveis.

Escopo: Local e global. Provê uma visão geral ou visualizações granulares de partes locais do conjunto de dados.

Compreensão: Média - É usado para confirmar os resultados modelados pela IA; para confirmar se a IA está classificando corretamente - i.e., quando se

conhece as hierarquias, classes ou cluster dos dados de treinamento ou teste -; para confirmar se atributos e estruturas similares são projetados relativamente perto e atributos dispares longe uns dos outros.

A.1.3 t-SNE

t-SNE é uma ferramenta para visualizar os dados de similaridade resultantes da conversão de um conjunto de dados de alta dimensão em uma matriz de similaridades. É capaz de capturar muito bem a estrutura local dos dados de alta dimensão, além de revelar a estrutura global, i.e., a presença de clusters em várias escalas (VAN DER MAATEN e HINTON, 2008).

Especificidade: Genérico - Pode ser usada para visualizar conjuntos de dados complexos com muitas variáveis.

Escopo: Local e global. Provê uma visão geral ou visualizações granulares de partes locais do conjunto de dados.

Compreensão: Média - É usado para confirmar os resultados modelados pela IA; para confirmar se a IA está classificando corretamente - i.e., quando se conhece as hierarquias, classes ou cluster dos dados de treinamento ou teste -; para confirmar se atributos e estruturas similares são projetados relativamente perto e atributos dispares longe uns dos outros.

A.1.4 Autoencoder

Autoencoder Networks são algoritmos de compactação de dados em que as funções de compactação e descompactação são: específicas dos dados (i.e., eles só poderão compactar dados semelhantes aos que foram treinados), com perdas (i.e., as saídas descomprimidas serão degradadas em comparação com as entradas originais) e aprendidas automaticamente a partir de exemplos (i.e., é fácil treinar instâncias especializadas do algoritmo que terão bom desempenho em um tipo especí-

fico de entrada). Duas aplicações práticas interessantes são a eliminação de ruído em dados e a redução de dimensionalidade para visualização de dados (CHOLLET, 2016).

Especificidade: Genérico - Pode ser usada para visualizar conjuntos de dados complexos com muitas variáveis.

Escopo: Local e global. Provê uma visão geral ou visualizações granulares de partes locais do conjunto de dados.

Compreensão: Média - É usado para confirmar os resultados modelados pela IA; para confirmar se a IA está classificando corretamente - i.e., quando se conhece as hierarquias, classes ou cluster dos dados de treinamento ou teste -; para confirmar se atributos e estruturas similares são projetados relativamente perto e atributos dispares longe uns dos outros.

A.1.5 Disentangled Inferred Prior Variational Autoencoder (DIP-VAE)

O algoritmo Disentangled Inferred Prior Variational Autoencoder (DIP-VAE) é um algoritmo não supervisionado de representação de aprendizado que recebe os atributos emaranhados, i.e., vários atributos significativos combinados num só, como entrada e aprende uma nova representação para eles, produzindo como saída uma nova representação desemaranhada e compreensível dos atributos (GURUMOORTHY, DHURANDHAR, *et al.*, 2019).

Especificidade: Genérico.

Escopo: Global.

Compreensão: Alta.

A.1.6 Grafo de Correlação

Um grafo de correlação é uma representação bidimensional dos relacionamentos (correlações) em um conjunto de dados. Eles nos permitem ver grupos de variáveis correlacionadas, identificar variáveis irrelevantes e descobrir ou verificar relacionamentos importantes que os modelos de aprendizado de máquina devem incorporar.

Especificidade: Genérico - Pode ser usada para visualizar conjuntos de dados complexos com muitas variáveis.

Escopo: Local e global - Provê uma visão geral ou visualizações granulares de partes locais do conjunto de dados.

Compreensão: Média - Evidenciam relacionamentos complexos e importantes nos dados. Relacionamentos esperados (e.g., o atributo salário com o resultado da previsão da sua pontuação de crédito) que ocorrem nos dados devem ser refletidos pela IA. Variáveis correlacionadas, por arestas mais grossas, com as saídas são importantes para o modelo e variáveis desconexas possuem pouca relevância.

A.2 EXPLICAÇÃO DO MODELO

Os modelos de aprendizagem de máquina após seu treinamento geralmente carecem de explicações quanto ao seu comportamento. Assim, é benéfico compreender as características e atributos desses modelos antes de implementá-los. As principais técnicas de explicação do modelo foram agrupadas em três categorias: explicação direta, explicação indireta e exemplos.

A.2.1 Explicação direta

Um critério importante para o sucesso de um projeto de IA é determinar o grau de interpretabilidade necessária. Se prover explicação é vital para o seu projeto, é melhor usar uma técnica de modelagem interpretável (i.e., devido à sua estrutura simples) desde o início.

As principais técnicas de explicação direta (ou modelos autoexplicáveis) são: árvore de decisão, modelos baseados em regras e regressão linear.

A.2.1.1 Árvore de decisão

As árvores de decisão são fluxogramas derivados dos dados nos quais cada nó interno representa um teste em um atributo; cada ramificação representa o resultado do teste; e cada folha representa um rótulo de classe. Os caminhos da raiz para a folha representam regras de classificação e podem ser visualizados ou explicados por regas SE-ENTÃO.

Especificidade: Específico – Dado que a interpretabilidade é o principal motivador para optar-se por usá-las.

Escopo: Global – Uma vez que nos ajudam a entender as relações funcionais do modelo e o relacionamento entre as previsões e as variáveis de entrada.

Compreensão: Alta - a estrutura em árvore induz-nos a pensar nas previsões para cada instância como contrafactuais. As explicações da árvore são contrastantes, dado que você sempre pode comparar uma previsão com os cenários alternativos relevantes, i.e., as outras folhas da árvore. Contudo, as árvores não conseguem lidar bem com relacionamentos lineares; são bastante instáveis, i.e., uma leve alteração nos dados de treinamento pode criar uma árvore completamente diferente; e são muito interpretáveis, desde que sejam curtas.

A.2.1.2 Modelos baseados em Regras

Uma regra de decisão é uma instrução SE-ENTÃO que consiste em uma estrutura geral: SE as condições forem atendidas ENTÃO faça uma certa previsão. Uma regra de decisão pode ser combinada com outras para gerarem previsões.

Há muitas maneiras de inferir regras a partir dos dados. Os principais algoritmos são: o OneR que aprende regras a partir de um único atributo. É caracterizado por sua simplicidade, interpretabilidade e seu uso como referência (HOLTE, 1993); o *Sequential covering* que aprende regras de forma iterativa e remove os pontos de dados cobertos pela nova regra. Este procedimento é o mais usado pelos algoritmos de aprendizagem de regras (COHEN, 1995); e o *Bayesian Rule Lists* que combinam padrões pré-minerados em uma lista de decisão usando estatística bayesiana (LETHAM, RUDIN, *et al.*, 2015).

Especificidade: Específico – Dado que a interpretabilidade é o principal motivador para optar-se por usá-las.

Escopo: Global – Uma vez que nos ajudam a entender as relações funcionais do modelo e o relacionamento entre as previsões e as variáveis de entrada.

Compreensão: Alta - Regras SE-ENTÃO são fáceis de interpretar, provavelmente o modelo mais interpretável, se o número de regras for pequeno; são resistentes contra transformações monotônicas dos atributos de entrada; podem ser tão expressivas quanto árvores de decisão, sendo mais compactas; mas são ruins para descrever relações lineares entre atributos e saída.

A.2.1.3 Regressão Linear

Modelos de regressão linear são utilizados por estatísticos, cientistas da computação e outros profissionais que lidam com problemas quantitativos. Um modelo de regressão linear calcula suas previsões como a soma ponderada dos atributos de entrada.

A maior vantagem dos modelos de regressão linear é a sua linearidade e, por conseguinte, o procedimento de estimativa é simplificado, i.e., equações lineares têm uma interpretação fácil de entender. Essa é uma das principais razões pela qual modelos lineares e similares são tão difundidos em áreas como medicina, sociologia, psicologia etc.

Especificidade: Específico - Na prática, devido as implementações de restrições de cada modelo.

Escopo: Global - As restrições de monotonicidade criam funções de resposta globalmente interpretáveis.

Compreensão: Baixa – As explicações são contrastantes, mas a instância de referência é um ponto de dados em que todos os recursos numéricos são zero e, geralmente, é um exemplo artificial, sem sentido, improvável de ocorrer nos dados ou na realidade.

A.2.2 Explicação Post hoc ou indireta

A interpretabilidade indireta refere-se à aplicação de métodos de interpretabilidade ao modelo treinado para extrair explicações sobre seu funcionamento. As principais técnicas de explicação indireta (ou modelos sucedâneos) são: Individual Conditional Expectation (ICE), Modelo Substituto Global, Modelo Substituto Local, Regras de Escopo (Anchors), Treeinterpreter e Valores Shapley.

A.2.2.1 Individual Conditional Expectation (ICE)

Os gráficos de dependência parcial (GDP) ajudam a visualizar o relacionamento parcial médio entre a previsão (i.e., saída) e um ou mais atributos (i.e., entradas). Contudo, quando há muita interação entre os atributos, o relacionamento parcial pode ser heterogêneo. Acarretando assim, num ofuscamento da complexidade do relacionamento modelado (GOLDSTEIN, KAPELNER, *et al.*, 2015).

O Individual Conditional Expectation (ICE) plots é uma ferramenta para visualizar um modelo estimado por qualquer algoritmo de aprendizado supervisionado. O ICE refina o GDP ao representar graficamente a relação funcional entre a previsão (i.e., resposta) e os atributos (i.e., entradas) para cada observação individual (GOLDSTEIN, KAPELNER, *et al.*, 2015).

O ICE mostra como um modelo se comporta para uma única linha de dados e pode ser usado para validar restrições de monotonicidade, i.e., sugerindo onde e em que extensão podem existir heterogeneidades (GOLDSTEIN, KAPELNER, *et al.*, 2015). Como o modelo pode detectar fortes interações entre as variáveis de entrada, o ICE combinado com o GDP fornece informações locais e expande o entendimento global fornecido pelo GDP.

Especificidade: Genérico– Podem ser aplicados a diferentes tipos de modelos.

Escopo: Local – Porque se aplica a uma observação de cada vez.

Compreensão: Média – Para cada observação, o ICE mostra as interações entre as entradas e saídas do modelo, além do comportamento, i.e., não linearidade e monotonicidade. Contudo, podem exibir apenas um atributo significativamente e nem sempre é fácil enxergar a média.

A.2.2.2 Modelo Substituto Global

Um modelo substituto global é um modelo interpretável, treinado para aproximar as previsões de um modelo caixa preta. Dessa forma, podemos tirar conclusões sobre o funcionamento do modelo caixa preta interpretando o modelo substituto.

A confecção de um modelo substituto global, cabendo pequenas variações, é descrito a seguir:

(1) Assumindo a existência do modelo caixa preta (M);

(2) Selecione um conjunto de dados (D), e.g., pode ser o mesmo conjunto de dados usado para treinar M (ou um subconjunto desses dados) ou um novo conjunto com a mesma distribuição;
(3) Para o conjunto de dados selecionado D, obtenha as previsões (P) do modelo caixa preta M;
(4) Isso gera um novo conjunto de dados (B), formado pelas entradas de D e as previsões do modelo caixa preta P;
(5) Escolha um modelo interpretável (I) (árvore de decisão etc.);
(6) Treine o modelo interpretável I no conjunto de dados B;
(7) Ao final do treinamento, I é um modelo substituto de P;
(8) Avalie a paridade entre I e P;
(9) Interprete I.

Uma maneira, das várias encontradas na literatura, de se avaliar a paridade entre os modelos substituto e caixa preta é por meio do coeficiente de determinação (R^2)[87].

Especificidade: Genérico– Uma vez que não requer nenhuma informação sobre o funcionamento do modelo caixa preta.

Escopo: Global – Uma vez que se trata de um mapeamento entre entradas e previsões.

Compreensão: Alta – A metodologia é flexível, i.e., permite usar modelos distintos para se adequar a audiência da explicação, a abordagem é muito intuitiva e direta e podemos medir quão bons são nossos modelos substitutos. Porém, deve-se tomar cuidado para suas explicações serem sobre o modelo caixa preta e não sobre os dados de treinamento; com a escolha do modelo substituto; com a falta de compatibilidade entre os modelos, uma vez que se trata de uma aproximação; além de não sabermos qual o valor ideal para o coeficiente de determinação.

[87] R^2 é uma medida de ajustamento de um modelo estatístico linear generalizado em relação aos valores observados. Varia entre 0 e 1, indicando, em percentagem, o quanto o modelo consegue explicar os valores observados. Quanto maior o R^2, mais explicativo é o modelo, melhor ele se ajusta à amostra.

A.2.2.3 Modelo Substituto Local

Modelos substitutos locais são modelos interpretáveis usados para explicar previsões individuais de modelos caixa preta. O Local Interpretable Model-Agnostic Explanations (LIME) é uma implementação concreta desse conceito.

A ideia por trás do LIME é bem intuitiva. Consiste em usar o modelo de caixa preta como uma espécie de oráculo a quem podemos indagar quantas vezes queiramos. Como nossa meta é entender o porquê de certa previsão, precisamos estabelecer o que e como perguntar ao nosso oráculo.

O LIME gera um novo conjunto de dados que consiste em amostras permutadas e as respectivas previsões do oráculo. A forma como geramos essas permutações varia de acordo com a natureza dos dados, e.g., para textos, a partir do texto original, novos textos são criados removendo aleatoriamente as palavras do texto original e para imagens, variações das imagens são criadas segmentando a imagem em "superpixels" e ativando ou desativando os superpixels (RIBEIRO, SINGH e GUESTRIN, 2016).

Também, varia a forma como amostramos a vizinhança, i.e., escolhemos os pontos de dados circunvizinhos para formar a amostra. Atualmente o LIME usa um kernel de suavização exponencial para definir a vizinhança, i.e., uma função que recebe dois pontos de dados e retorna uma medida de proximidade (RIBEIRO, SINGH e GUESTRIN, 2016).

O passo seguinte é treinar um modelo interpretável, e.g., árvore de decisão, como os dados arrolados. O modelo resultante é uma boa aproximação local do modelo caixa preta, mas não necessariamente uma boa aproximação global.

Essa aproximação é denominada fidelidade. A fidelidade nos dá uma boa ideia do quão confiável é o modelo interpretável para explicar as previsões da caixa preta na vizinhança da instância de dados de interesse.

Especificidade: Genérico– Podem ser aplicados a diferentes tipos de modelos.

Escopo: Local – Nos ajuda a entender as previsões do modelo para um único ponto de dados ou um grupo semelhante.

Compreensão: Alta – Fácil de usar, funciona com dados tabulados, texto e imagens, as explicações podem ser curtas e contrastantes e os dados coletados podem ser usados para gerar vários modelos substitutos diferentes. Todavia, a definição correta de vizinhança é um problema não resolvido para algumas aplicações.

A.2.2.4 Regras de Escopo (Anchors)

Anchors explica previsões individuais de qualquer modelo caixa preta, encontrando uma regra de decisão que "ancora" a previsão. Uma regra "ancora" uma previsão se alterações nos demais atributos não afetarem a previsão. O Anchors utiliza técnicas de aprendizado por reforço em combinação com um algoritmo de pesquisa em grafos para reduzir o número de acessos ao modelo ao mínimo, enquanto ainda livrando-se dos ótimos locais (RIBEIRO, SINGH e GUESTRIN, 2018).

O Anchors usa quatro componentes principais para encontrar explicações:

(1) Geração de candidatos: gera novos candidatos a explicação;

(2) Identificação do melhor candidato: as regras dos candidatos devem ser comparadas em relação à regra que melhor explica;

(3) Validação de precisão do candidato: coleta mais amostras caso ainda não haja confiança estatística de que o candidato exceda o limiar de precisão estabelecido;

(4) Beam Search modificada: Todos os componentes acima são montados em uma beam search – i.e., um algoritmo de pesquisa heurística que explora um grafo expandindo o nó mais promissor em um conjunto limitado – que carrega os M melhores candidatos para a próxima rodada. Quando M excede o limiar de precisão, a busca retorna o candidato que tem a melhor cobertura (i.e., a "ancora") (RIBEIRO, SINGH e GUESTRIN, 2018).

Especificidade: Genérico– Podem ser aplicados a diferentes tipos de modelos.

Escopo: Local – Nos ajuda a entender as previsões do modelo para um único ponto de dados ou um grupo semelhante.

Compreensão: Alta – Regras são fáceis de interpretar, o algoritmo pode ser paralelizado e funciona quando as previsões do modelo são não-lineares ou complexas. Contudo, requer muitos acessos ao modelo, a configuração é complexa e.g., largura da busca, limiar de precisão, e a cobertura é indefinida em alguns domínios.

A.2.2.5 Treeinterpreter

O *Treeinterpreter* decompõe as previsões da árvore de decisão, e outros algoritmos baseados em árvore, em influência (média geral dos dados de treinamento) e termos de componente para cada variável usada em um modelo. O método simplesmente gera uma lista das influências e contribuições globais de cada variável.

Especificidade: Genérico – Específico para algoritmos baseados em árvores de decisão.

Escopo: Global e Local – Global quando representa as contribuições médias dos atributos (i.e., entradas) para as previsões (i.e., saídas) do modelo e local quando usada para explicar uma única previsão.

Compreensão: Média – Aumenta a compreensão exibindo as contribuições das variáveis de entrada para as previsões. No entanto, as contribuições locais não condizem com a previsão do modelo em alguns casos e em algumas implementações.

A.2.2.6 Valores Shapley

Explicações Shapely derivam as contribuições dos atributos nas previsões dos modelos caixa preta. Ao modelarmos uma previsão (i.e., a saída do modelo) como um montante a ser distribuído entre os atributos (i.e., entradas do modelo) podemos estimar qual a contribuição de cada um dos atributos para a previsão. Os valores Shapley nos dizem como atribuir de maneira "justa" a previsão entre os atributos.

Por atribuição de maneira justa, subentende-se satisfazer as seguintes propriedades: eficiência, simetria, neutralidade e linearidade.

Eficiência - A soma dos Shapleys de todos os atributos é igual ao valor da previsão;

Simetria - As contribuições de dois atributos j e k devem ser iguais se contribuírem igualmente para todas as previsões;

Neutralidade - Um atributo j que não altera o valor da previsão tem um Shapley igual a 0;

Linearidade - As contribuições de dois atributos j e k é igual à soma dos seus Shapleys.

Explicações Shapely são baseadas em contribuições locais precisas das variáveis de entrada e podem ser ordenadas para gerar interpretações automáticas. Por serem embasadas na teoria dos jogos, podem ser usadas quando houver necessidades legais de explicabilidade, mas são computacionalmente custosas para se obter.

Especificidade: Genérico– Podem ser aplicados a diferentes tipos de modelos.

Escopo: Local e Global – As explicações são locais, mas podem ser agregadas para criar explicações globais.

Compreensão: Alta – Permite explicações contrastivas, a diferença entre a previsão e a média é razoavelmente distribuída entre os valores dos atributos

e é o único método de explicação com uma teoria sólida. Não obstante, são computacionalmente custosos para calcular, você precisa ter acesso aos dados para calcular o valor de uma nova instância, as explicações sempre incluem todos os atributos e podem ser mal interpretadas, i.e., o Shapley de um atributo é decorrente dos demais existentes.

A.2.3 Exemplos

Os métodos de explicação baseados em exemplos selecionam instâncias específicas dos dados para explicar o comportamento dos modelos ou para explicar a distribuição de dados subjacente. Os principais métodos de interpretação baseados em exemplo são: contrafactuais, protótipos e críticas (MMD-critic e ProtoDash), instâncias influentes, diagnóstico por exclusão e funções de influência.

A.2.3.1 Explicações Contrafactuais

Contrafactuais requerem imaginar uma realidade hipotética que contradiz os fatos observados, daí o nome “contrafactual”. Uma explicação contrafactual descreve uma situação causal da forma: “Se X não tivesse ocorrido, Y não teria ocorrido”.

Dessa forma, o contrafactual de uma previsão descreve a menor alteração nos atributos (i.e., entradas) que modificam a previsão (i.e., saída) para um resultado desejado (WACHTER, MITTELSTADT e RUSSELL, 2018). Diferentemente dos protótipos, os contrafactuais não precisam ser instâncias reais dos dados, mas podem ser uma nova combinação de atributos.

Especificidade: Genérico– Uma vez que o método funciona apenas com as entradas e saídas do modelo.

Escopo: Local – Contrafactuais aumentam a compreensão criando explicações para uma única previsão.

Compreensão: Alta – A interpretação das explicações contrafactuais é direta, i.e., se os atributos de uma instância forem alterados de acordo com o contrafactual, a nova previsão será a previsão estipulada. É relativamente fácil de implementar e não requer acesso aos dados ou ao modelo, i.e., apenas acesso à função de previsão do modelo. Isso é atraente para empresas auditadas por terceiros ou que oferecem explicações para os usuários sem divulgar o modelo ou os dados.

A.2.3.2 Protótipos e Críticas

Um protótipo é uma amostra representativa dos dados. Uma crítica é uma amostra dos dados que não é bem representada pelo conjunto de protótipos. Protótipos e críticas podem ser usados independentemente para descrever os dados, criar um modelo interpretável ou para interpretar um modelo de caixa preta. Veremos duas abordagens que combinam protótipos e críticas em uma única estrutura.

A.2.3.2.1 MMD-critic

O MMD-critic compara a distribuição dos dados e a distribuição dos protótipos selecionados. Então, ele seleciona protótipos que minimizam a discrepância entre as duas distribuições. Os pontos de dados em áreas com alta densidade são bons protótipos, especialmente quando os pontos são selecionados em diferentes clusters. Os pontos de dados de regiões que não são bem explicadas pelos protótipos são selecionados como críticas (KIM, KHANNA e KOYEJO, 2016).

Especificidade: Genérico – Funciona com qualquer tipo de dados e qualquer tipo de modelo de aprendizado de máquina.

Escopo: Global – Você pode usar o MMD-critical para encontrar protótipos e críticas nos dados e em seguida usar o modelo treinado para prever os resultados para os protótipos e críticas. Da análise das previsões, você terá vários

exemplos que representam bem os dados e exemplos que evidenciam os pontos fracos do modelo treinado.

Compreensão: Média – Apesar de fornecer evidências de pontos positivos e negativos do seu modelo, o método não explica o porquê deles. Além disso, como a distinção entre protótipos e críticas é baseada em um valor de corte (i.e., número de protótipos), as críticas podem acabar em áreas que não são tão bem explicadas.

A.2.3.2.2 ProtoDash

O algoritmo ProtoDash além de selecionar protótipos, também associa pesos não negativos, como indicativo de sua importância. Esta extensão fornece uma estrutura unificada sob a qual protótipos e críticas (ou seja, outliers) podem ser encontrados. (KUMAR, SATTIGERI e AVINASH, 2018).

Especificidade: Genérico - Funciona com qualquer tipo de dados e qualquer tipo de modelo de aprendizado de máquina

Escopo: Global - Você pode usar o PhotoDash para encontrar protótipos e críticas nos dados e sem seguida usar o modelo treinado para prever os resultados para os protótipos e críticas. Da análise das previsões, você terá vários exemplos que representam bem os dados e exemplos que evidenciam os pontos fracos do modelo treinado.

Compreensão: Médio - Apesar de fornecer evidências de pontos positivos e negativos do seu modelo, o método não explica o porquê deles.

A.2.3.3 Instâncias Influentes

Em última análise, um modelo treinado de *machine learning* é fruto dos dados de treinamento e, a exclusão de uma das instâncias de treinamento pode afetar o modelo final. Uma instância é dita influente quando sua remoção dos dados de trei-

namento tem um forte impacto no modelo treinado, i.e., quanto maior a mudança nos parâmetros e previsões do modelo, maior a influência da instância.

Ao identificar instâncias influentes para o treinamento, podemos depurar modelos de aprendizado de máquina e explicar melhor seus comportamentos e previsões. O fito por trás das instâncias acolitarem a interpretabilidade, é rastrear os parâmetros e previsões do modelo de volta aos dados de treinamento. Veremos duas abordagens que identificam instâncias influentes, diagnóstico por exclusão e funções de influência.

A.2.3.4 Diagnóstico por exclusão

No diagnóstico por exclusão, nós elidimos a instância que queremos analisar dos dados de treinamento; retreinamos o modelo no novo conjunto de dados (i.e., sem a instância que estamos analisando); e comparamos as diferenças nos parâmetros e previsões do modelo (individualmente ou nos dados completos). Podemos remover a instância a ser estudada dos parâmetros ou das previsões do modelo.

O DFBETA mede o efeito de excluir uma instância nos parâmetros do modelo.

$$\text{DFBET}\, A_i = \beta - \beta^{(-i)}$$

Onde β é o vetor de ponderação quando o modelo é treinado em todas as instâncias de dados e $\beta^{(-i)}$ o vetor de ponderação quando o modelo é treinado sem a i-ésima instância (BELSLEY, KUH e WELSH, 1980).

A distância de Cook mede o efeito de excluir uma instância nas previsões do modelo, i.e., o quanto a saída de um modelo linear muda quando removemos a i-ésima instância do treinamento.

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j \ - \hat{y}_j^{(-i)} \right)^2}{p \cdot EMQ}$$

Onde o numerador é a diferença quadrática entre a previsão do modelo com e sem a i-ésima instância, somada ao conjunto de dados. E o denominador é o número de atributos *p* vezes o erro médio quadrático (EMQ). (COOK, 1977).

Para ambas as medidas, precisamos retreinar o modelo, omitindo instâncias individuais a cada vez. Em seguida, os parâmetros ou previsões – do modelo original e do modelo com uma das instâncias excluídas dos dados de treinamento – são comparados.

Especificidade: Específico – DFBETA só funciona em modelos parametrizados, e.g., redes neurais, e a distância de Cook, apesar de ser global, usita o erro quadrático médio, que não é significativo para todos os modelos preditivos, e.g., classificadores.

Escopo: Global e Local – Na forma como foram apresentadas, elas são globais, mas nada impede que sejam generalizadas para análises locais.

Compreensão: Alta – Estão entre as melhores ferramentas de depuração para modelos de aprendizado de máquina que vimos até aqui. Nos permitem comparar diferentes modelos de aprendizado de máquina, mas os diagnósticos de exclusão são muito caros (computacionalmente) porque exigem retreino.

A.2.3.5 Funções de Influência

As funções de influência ajudam a entender o comportamento do modelo, depurar o modelo e detectar erros no conjunto de dados. Elas mapeiam os parâmetros e previsões do modelo de volta à instância de treinamento responsável, i.e., como se fosse uma função inversa à do modelo.

Em vez de excluir instâncias como nos métodos anteriores, o preceito por trás das funções de influência é aumentar a perda de uma instância de treinamento por um fator infinitesimalmente pequeno ε, o que resulta em novos parâmetros no modelo. O método requer, portanto, acesso ao gradiente de perda com relação aos parâ-

metros do modelo, o que limita sua aplicação a um subconjunto de modelos, e.g., redes neurais e SVM são contempladas, mas árvores não.

As funções de influência têm muitas aplicações. Entender o comportamento (i.e., entender o funcionamento e as fraquezas) do modelo analisado e com isso extrair um entendimento de como ele "pensa"; corrigir dados de treinamento, e.g., escolher as instâncias mais influentes para corrigir primeiro; e sanar incompatibilidade de domínios – i.e., a distribuição dos dados de treinamento e teste são diferentes, o que pode causar um desempenho ruim do modelo nos dados de teste – técnica muito usada para elicitar problemas de transferência de conhecimento.

Especificidade: Específico – Mas podem ser aplicadas a uma ampla classe de modelos.

Escopo: Global e Local – Na forma como foram apresentadas, elas são globais, mas nada impede que sejam generalizadas para análises locais.

Compreensão: Alta – As funções de influência, por meio de derivativos, podem ser usadas para criar dados de treinamento antagônicos (i.e., adversariais). São uma boa alternativa para o diagnóstico por exclusão, mas apenas para modelos com parâmetros diferenciáveis e não tratam de possíveis correlações entre instâncias.

A.3 IMPARCIALIDADE

As principais técnicas de mitigação de viés são: Pré-processamento Otimizado, Pós-processamento de Probabilidades Equalizadas, Repesagem, Regularizador Removedor de Preconceitos, Gerador de Representações Justas e *Adversarial Debiasing*.

A.3.1 Pré-processamento Otimizado

A não discriminação é uma característica desejada nas previsões tomadas por inteligências artificiais. O método consiste em uma estrutura de otimização flexível para transformar dados probabilisticamente, a fim de reduzir a discriminação algorítmica, i.e., controlar a discriminação, limitar a distorção em amostras individuais e preservar a utilidade (CALMON, WEI, *et al.*, 2017).

Quando usado para treinar classificadores padrão, o conjunto de dados transformado levou a uma classificação mais justa, quando comparado ao conjunto de dados original. A redução da discriminação tem uma penalidade de precisão devido às restrições impostas ao mapeamento aleatório.

Além disso, o método é competitivo com outros na literatura, com o benefício adicional de permitir um controle explícito da justiça individual e a possibilidade de variáveis protegidas multivaloradas e não binárias.

População: Individual e grupo.

Fase: Pré-treinamento.

Visão: Isonomia.

Metas: controlar a discriminação, limitar a distorção em amostras individuais e preservar a utilidade.

A.3.2 Pós-Processamento de Probabilidades Equalizadas

Probabilidades equalizadas são uma medida de justiça que realiza duas pretensões importantes. Primeiro, elas corrigem as principais deficiências conceituais da paridade demográfica como uma noção de justiça. E em segundo lugar, estão totalmente alinhadas com o objetivo de construir classificadores mais precisos (HARDT, PRICE e SREBRO, 2016).

Probabilidades equalizadas criam uma estrutura de incentivos para os fabricantes construírem preditores alinhados às metas de justiça. No entanto, para se alcançar melhores previsões com probabilidades equalizadas, precisamos coletar atributos que capturam mais diretamente o alvo, não relacionados à sua correlação com o atributo protegido (HARDT, PRICE e SREBRO, 2016).

O preditor equalizado de probabilidade, usado na fase de pós-processamento, depende do mínimo da curva ROC entre os diferentes grupos protegidos, incentivando a construção de preditores precisos para todos os grupos. Essa etapa requer apenas informações agregadas sobre os dados e pode até ser realizada usando privacidade diferencial, por exemplo.

População: Grupo.

Fase: Pós-treinamento

Visão: Isonomia.

Metas: Corrigir as principais deficiências conceituais da paridade demográfica como uma noção de justiça e construir classificadores mais precisos e justos.

A.3.3 Repesagem

Essa técnica nos apresenta três abordagens para o problema da classificação com restrições de não discriminação, todos baseados no pré-processamento dos dados de treinamento. São elas: reetiquetagem, repesagem e amostragem.

A primeira abordagem, denominada reetiquetagem dos dados, baseia-se na alteração dos rótulos das classes para remover a discriminação dos dados de treinamento. A segunda abordagem, repesagem, é menos invasiva, dado que não altera os rótulos das classes. Em vez disso, pesos são atribuídos aos dados para torná-los livre de discriminação. Na terceira, amostragem, extraímos novamente do conjunto de dados de forma que a discriminação seja removida (KAMIRAN e CALDERS, 2012).

Todas as soluções apresentadas são baseadas na remoção da discriminação dos dados de treinamento com mais eficiência que métodos simples, e.g., remover o atributo sensível dos dados de treinamento. Posteriormente, um classificador é treinado nestes dados saneados. O juízo por trás dessa abordagem é que, como o classificador é treinado em dados livres de discriminação, é provável que suas previsões sejam (mais) livres de discriminação.

População: Grupo.

Fase: Pré-treinamento

Visão: Equidade

Metas: Fornecer a melhor estratégia entre precisão e não discriminação para classificadores.

A.3.4 Regularizador removedor de preconceitos

A técnica é usada para atenuar viés em classificadores. Adiciona um termo de regularização que reconhece discriminação aos objetivos de aprendizado. Segundo essa técnica, existem três causas, i.e., tipos de preconceito, de injustiça no aprendizado de máquina. Preconceito significa uma dependência estatística entre uma variável sensível, S, e a variável de destino, Y, ou uma variável não sensível, X. Existem três tipos de preconceitos: preconceito direto, preconceito indireto e preconceito latente.

O primeiro tipo é o preconceito direto, que é o uso de uma variável sensível em um modelo de previsão. O segundo tipo é o preconceito indireto, que é a dependência estatística entre uma variável sensível e uma variável de destino. O terceiro tipo de preconceito é o preconceito latente, que é uma dependência estatística entre uma variável sensível, S e uma variável não sensível, X (KAMISHIMA, AKAHO, *et al.*, 2012).

Definidas as métricas, o método propõe uma abordagem de regularização aplicável a qualquer algoritmo de previsão com modelo probabilístico discriminativo.

Objetivando a classificação, dois regularizadores, baseados em modelos de regressão logística, são construídos.

O primeiro regularizador é padrão para evitar *over-fitting*[88]. O segundo regularizador, i.e., removedor de preconceito, é introduzido para impor uma classificação justa, i.e., tenta diretamente reduzir o índice de preconceito. Esses reguladores são acrescentados aos objetivos de aprendizado, i.e., eles aumentam uma função de perda padrão com um regularizador de justiça.

População: Individual e grupo.

Fase: In-treinamento

Visão: Equidade ou Isonomia.

Metas: Atenuar viés em classificadores adiciona um termo de regularização que reconhece discriminação aos objetivos de aprendizado.

A.3.5 Gerador de Representações Justas

O método formula a equidade como um problema de otimização para encontrar uma representação intermediária dos dados – que pode ser usada para outras tarefas de classificação, e.g., transferência de aprendizado – que melhor os codifique, enquanto ofusca aspectos deles, removendo simultaneamente qualquer informação sobre associação com grupo protegido.

O modelo mapeia cada indivíduo, representado como um ponto de dados em um determinado espaço de entradas, para uma distribuição de probabilidade em um novo espaço de representação. O objetivo da nova representação é "perder" qualquer informação que possa estabelecer se o indivíduo pertence a subgrupos protegidos, mantendo o máximo possível de outras informações (ZEMEL, WU, *et al.*, 2013).

---

[88] quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados.

A justiça se torna um problema de otimização para encontrar a representação intermediária que melhor codifica os dados, ofuscando a associação aos subgrupos protegidos. As classificações podem ser feitas com base nessas novas representações.

População: Individual e grupo.

Fase: In-treinamento

Visão: Equidade.

Metas: Classificação justa por meio da melhor codificação possível dos dados e ofuscação de qualquer informação de pertinência ao grupo protegido.

A.3.6 Adversarial Debiasing

É um método geral e poderoso para o treinamento de modelos imparciais de aprendizado de máquina. O método visa mitigar vieses, incluindo uma variável para o grupo de interesse e treinando simultaneamente um preditor e um adversário.

São abordadas três medidas de equidade: paridade demográfica, igualdade de probabilidades e igualdade de oportunidades. As medidas de equidade são tratadas no contexto de *adversarial debiasing*. Nesse contexto, a tarefa é prever uma variável de saída Y, dada uma variável de entrada X, mantendo-se imparcial em relação a alguma variável Z, i.e., variável protegida. Para esses sistemas de aprendizagem, o preditor $\hat{Y} = f(X)$ pode ser construído como triplas (X, Y, Z), i.e., (entrada, saída, variável protegida) (ZHANG, LEMOINE e MITCHELL, 2018).

O preditor $f(X)$ geralmente recebe acesso à variável protegida Z, embora isso não seja estritamente necessário. Essa construção permite determinar quais tipos de vieses são considerados indesejáveis para que uma aplicação específica seja escolhida através da especificação da variável protegida.

Em seguida, um preditor f será treinado para modelar Y com a maior precisão possível, satisfazendo as restrições de igualdade. A paridade demográfica será alcançada com a introdução de um adversário g, que tentará prever um valor para Z a partir de $\hat{Y}$. O gradiente de g será incorporado à regra de atualização de peso de f, a fim de reduzir a quantidade de informações sobre Z transmitidas através de $\hat{Y}$ (ZHANG, LEMOINE e MITCHELL, 2018).

População: Individual e grupo.

Fase: In-treinamento.

Visão: Equidade e isonomia.

Metas: Mitigar o viés nas três medidas de equidade abordadas: paridade demográfica, igualdade de probabilidades e igualdade de oportunidades.