Ética e Segurança da Inteligência Artificial

Ferramentas práticas para se criar "bons" modelos

Nicholas Kluge Corrêa[1]

AIRES PUCRS

[1] Mestre em Engenharia Elétrica e Doutorando em Filosofia – PUCRS. Bolsista do Programa de Excelência Acadêmica (Proex) da Fundação CAPES (Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior). Presidente da AIRES PUCRS.

Prefácio

> “Someday a computer will give a wrong answer to spare someone's feelings, and man will have invented artificial intelligence”.
>
> — Robert Breault

A **AI Robotics Ethics Society (AIRES)** é uma organização sem fins lucrativos fundada em 2018 por Aaron Hui, com o objetivo de se promover a conscientização e a importância da implementação e regulamentação ética da AI.

A AIRES é hoje uma organização com capítulos em universidade como UCLA (Los Angeles), USC (University of Southern California), Caltech (California Institute of Technology), Stanford University, Cornell University, Brown University e a Pontifícia Universidade Católica do Rio Grande do Sul (Brasil).

AIRES na PUCRS é o primeiro capítulo internacional da AIRES, e como tal, estamos empenhados em promover e aprimorar a Missão AIRES. Nossa missão é focar na educação dos líderes da IA de amanhã em princípios éticos, para assegurar que a IA seja criada de forma ética e responsável.

Como ainda existem poucas propostas de como devemos implementar princípios éticos e diretrizes normativas na prática do desenvolvimento de sistemas de IA, o objetivo deste trabalho é tentar diminuir esta lacuna entre o discurso e a práxis. Entre princípios abstratos e implementação técnica. Neste trabalho, buscamos introduzir o leitor ao tema da Ética e Segurança da IA. Ao mesmo tempo, apresentamos uma série de ferramentas para auxiliar desenvolvedores de sistemas inteligentes a desenvolver “bons” modelos. Este trabalho é um guia em desenvolvimento publicado em inglês e português. Contribuições e sugestões são bem vindas.

Introdução

"By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it".

— Eliezer Yudkowsky

Não é uma situação incomum quando um indivíduo, ou um grupo de indivíduos, se encontra em frente a um decisor responsável por realizar alguma forma de julgamento com base em um conjunto de fatos e características observáveis (e.g., um Juiz em um tribunal civil, um avaliador em uma entrevista de emprego, ou um gerente de banco responsável por autorizar, ou não, um empréstimo). Contudo, o que é novo é a utilização de modelos de inferência estatística para automatizar tais processos (e.g., modelos criados por aprendizagem de máquina).

Conforme sistemas autônomos afetam cada vez mais pessoas e a sociedade, o entendimento dos potenciais riscos relacionados a tais sistemas (e como mitigá-los) deve ser aprofundado. Para antecipar, prevenir e mitigar as consequências indesejáveis de tais sistemas, é fundamental que compreendamos quando e como problemas podem ser introduzidos ao longo do ciclo de vida (i.e., coleta de dados, treinamento, validação, testagem, implantação, etc.) de tais sistemas.

Certamente que não podemos reduzir todos os tipos de sistemas inteligentes, ou "Inteligência Artificial" (IA), a apenas Aprendizagem de Máquina. Possuímos também a abordagem simbólica (Newell, 1990), o conexionismo (Churchland & Sejnowski, 1992), as metodologias híbridas (simbólica/conexionista), a abordagem matemático-universal (Hutter, 2005), entre diversas outras metodologias que buscam desenvolver sistemas capazes de simular certas capacidades cognitivas para resolver diversos tipos de problemas (e.g., algoritmos genéticos, programação dinâmica, agentes BDI, etc.).

Contudo, como aprendizagem de máquina é atualmente uma das metodologias mais amplamente adotada e utilizada para diversas aplicações, especialmente aprendizagem profunda com suas diferentes técnicas (e.g., aprendizagem supervisionada, semi-supervisionada, não-supervisionada, auto-supervisionada), iremos nos focar neste guia principalmente nos problemas que devemos enfrentar quando desenvolvendo aplicações que utilizam dessa metodologia.

Problemas e efeitos colaterais que surgem de técnicas como aprendizagem por reforço (Amodei et al., 2016), e riscos relacionados a potenciais sistemas de IA avançada criados por aprendizagem de máquina (Hubinger et al., 2019), não serão abordados neste guia.

Por mais que a aprendizagem por reforço ainda não tenha "alcançado o mainstream", ela definitivamente é uma metodologia capaz de gerar soluções inteligentes, sendo o paradigma mais próximo do que podemos vir a chamar de "IA genuína" ou "Inteligência Artificial Geral" (IAG).[2]

Ainda são necessários certos avanços para tornar aprendizagem por reforço o novo paradigma de soluções em aprendizagem de máquina (e.g., métodos eficientes para se desenvolver funções de recompensa ou melhorar eficiência de amostragem). Avanços esses que, de forma constante e progressiva, estão sendo alcançados (Ye et al., 2021). Contudo, é muito possível que, em um futuro próximo, problemas relacionados à aprendizagem por reforço (e.g., hackeamento de recompensa, exploração segura, corrigibilidade) venham a se manifestar em aplicações desenvolvidas para o mundo real.

Contudo, tais problemas não serão o foco deste documento. Aqui, adotaremos uma visão de "curto-prazo" para problemas envolvendo ética e segurança da IA, i.e., problemas que enfrentamos na atualidade, com os sistemas que possuímos e utilizamos.

[2] Inclusive, para pesquisadores como Silver et al. (2021), o objetivo genérico de se maximizar recompensa pode ser o suficiente para produzir a maioria dos comportamentos inteligentes estudados em inteligência artificial e natural.

Sistemas criados por aprendizagem de máquina (especificamente aprendizagem supervisionada) aprendem modelos de inferência estatística com base em conjunto de dados observados, com o intuito de generalizar suas classificações/previsões/decisões para novos dados. Contudo, muitas vezes esses sistemas podem criar modelos que carregam diversos tipos de vieses, ou até mesmo agir de forma indesejada:

- Sistemas de reconhecimento facial podem apresentar vieses de classificação racistas (Lohr, 2018; Nunes, 2019);
- Sistemas de NLP (Natural Language Processing) podem apresentar vieses sexistas e misóginos (Wolf et al., 2017; Balch, 2020);
- Sistemas de classificação podem discriminar membros da comunidade LGBTQ+ (Wang & Kosinski, 2017; Agüera y Arcas et al., 2018).

Exploremos mais afundo o exemplo que diz respeito a discriminação racial: em outubro de 2019, o então vigente ministro da Justiça Sérgio Moro apresentou a portaria Nº793 como uma forma de modernização das forças policiais brasileiras. Nunes (2019) aponta que desde a implementação de tais sistemas, a população negra tem sido desproporcionalmente afetada. Em 2019, 90.5% dos indivíduos presos flagrados por sistemas de reconhecimento facial e videomonitoramento eram negros, o estado da Bahia liderando o número de prisões através dessas novas tecnologias (51,7%), seguido pelo Rio de Janeiro (31,7%), Santa Catarina (7,3%), Paraíba (3,3%) e Ceará (0,7%).

De acordo com um relatório disponibilizado pela Coordenadoria de Defesa Criminal e pela Diretoria de Estudos e Pesquisas de Acesso à Justiça da Defensoria Pública do Rio de Janeiro,[3] entre 1º de junho de 2019 e 10 de

[3] Defensoria Pública do Estado do Rio de Janeiro. Disponível em: http://www.defensoria.rj.def.br/uploads/imagens/d12a8206c9044a3e92716341a99b2f6f.pdf.

março de 2020 houveram pelo menos 58 casos de reconhecimento de imagem equivocados, resultando em acusações injustas e até mesmo na prisão de indivíduos inocentes. De todos os acusados injustamente, 70% eram negros.

Mas por que isso ocorre?

Uma resposta simplista seria "A resposta está nos dados. Os dados que usamos são enviesados". Contudo, uma resposta mais verídica seria "É um problema complexo".

Existe muito que ainda não entendemos sobre tais sistemas. Na conferência NIPS de 2017 (Conference on Neural Information Processing Systems), Ali Rahimi[4] levantou um importante ponto sobre o estado atual do campo de pesquisa em Aprendizagem de Máquina: "*a aprendizagem de máquinas se tornou alquimia*". "Antigamente" (i.e., antes de aprendizagem profunda se tornar o paradigma), técnicas como regressão linear, regressão logística, máquina de vetores de suporte, garantiam soluções eficientes e interpretáveis. Contudo, não fomos capazes de utilizar tais metodologias para com problemas mais complexos (e.g., visão computacional).

Depois dos anos 2000, com a criação de processadores mais rápidos, GPUs capazes de otimizar treinamento de redes neurais profundas, e melhores algoritmos de aprendizagem e otimização, aprendizagem profunda finalmente se estabeleceu como o novo paradigma.

Contudo, é importante lembrarmos que, por mais que aprendizagem de máquina gere modelos de inferência estatística, aprendizagem de máquina *não é como estatística.* Em aprendizagem de máquina, nós possuímos muitas hipóteses e poucos teoremas. Por isso aprendizagem de máquina é mais próxima de uma disciplina como engenharia do que matemática. Nós descobrimos o que funciona por tentativa e erro. Nós

[4] Transcrição disponível em: https://www.zachpfeffer.com/single-post/2018/12/04/transcript-of-ali-rahimi-nips-2017-test-of-time-award-presentation-speech.

ainda não possuímos teoremas que nos permitam verificar de forma rigorosa o comportamento de tais sistemas, e assim, fazer previsões de como eles poderão vir a se comportar no futuro (OOD - “out-of-distribution”).

Nas palavras de Ali Rahimi:

> A alquimia não é ruim. Há um lugar para a alquimia. A alquimia “funcionava”. Os alquimistas inventaram a metalurgia, formas de tingir tecidos, nossos modernos processos de fabricação de vidro e medicamentos. Os alquimistas também acreditavam que poderiam curar doenças com sanguessugas e transmutar metais de base em ouro. Para que a física e a química do século XVII pudessem provocar uma mudança em nossa compreensão do universo (que agora experimentamos), os cientistas tiveram que desmontar 2.000 anos de teorias alquímicas. Se você está usando alquimia para criar um sistema de compartilhamento de fotos, tudo bem. Mas estamos além disso agora. Estamos construindo sistemas que governam a saúde e mediam nosso diálogo cívico. Sistemas que influenciam nossas eleições. Eu gostaria de viver em uma sociedade cujos sistemas são construídos em cima de um conhecimento verificável, rigoroso, profundo, e não sobre alquimia.

Em outras palavras, aprendizagem de máquina ainda precisa de mais estudo teórico. Contudo, não é claro que a indústria irá abrandar o seu progresso e desenvolvimento prático em nome da cautela e formalização das teorias que fundamentam a criação de seus produtos. E isso gera problemas.

Assim, acreditamos que se torna necessária a criação e formalização de um novo agente para operar dentro de organizações e empresas voltadas ao desenvolvimento de tecnologias e soluções que utilizem tais tipos de sistemas. Precisamos de engenheiros de segurança e eticistas especializados em aprendizagem de máquina, i.e., agentes responsáveis

por prevenir e mitigar os possíveis efeitos colaterais de sistemas criados por aprendizagem de máquina).

Tal ator seria responsável por auxiliar a implementar medidas de segurança durante todo o ciclo de vida de tais sistemas, de modo a garantir que certos princípios éticos sejam respeitados e implementados durante o desenvolvimento, implantação e monitoramento de tais sistemas.

Para suprir tal necessidade, uma das respostas propostas pela comunidade envolvida na área de Ética da Inteligência Artificial, como instituições governamentais, corporações privadas, instituições acadêmicas, sociedades civis, associações de profissionais e ONGs, vêm sendo a publicação de diversos mecanismos de governança principiológicos. Estes mecanismos podem ser definidos como códigos de ética, diretrizes, entre outros instrumentos de governança similares, ou seja, documentos normativos baseados em princípios éticos (Russell et al., 2015; Boddington, 2017; Goldsmith & Burton, 2017; Floridi et al., 2018; Greene et al., 2019).

A Ética da IA, por mais que seja um campo da Ética relativamente novo,[5] já possui literatura o bastante para que meta-análises do campo tenham sido realizadas (Jobin et al., 2019; Hagendorff, 2020; Fjeld et al., 2020). Essas metanálises apontam que existe uma convergência para um certo grupo de princípios éticos (valores) comumente defendidos:

Valores	**Descrição**
Transparência	Este princípio aponta um dos maiores défices nas técnicas contemporâneas de Aprendizagem de Máquina. Enquanto seres humanos esperam explicações que possam compreender, algoritmos de aprendizagem de máquina operam com base em cálculos estatísticos complexos que desafiam

[5] De acordo com Jobin et al. (2019), menos de 20% de todos os documentos sobre Ética da IA revisados em sua metanálise (84) têm mais do que quatro anos de idade. De acordo com a ONG AlgorithmWatch (2020), seu Inventário Global das Diretrizes Éticas da IA contém 173 documentos. Nenhum desses documentos é anterior ao ano de 2013. Desses documentos, apenas 2 tem sua origem atrelada ao Sul da África e Sul da Ásia (nenhum documento produzido pela América Latina é listado).

	traduções simples, tornando-os "opacos" (Mittelstadt et al., 2019).
Justiça/Equidade	As questões de justiça incluem problemas de igualdade de tratamento e distribuição justa de benefícios. Este princípio é geralmente trabalhado na literatura através de definições algorítmicas de justiça e equidade (e.g., Paridade estatística/demográfica, Paridade Preditiva, Probabilidades Equalizadas) (Galhotra et al., 2017; Verma & Rubin, 2018).
Privacidade	Dados são como carvão para a indústria da IA. E as grandes empresas de tecnologia (Google, Amazon, Facebook), são as novas "minas de carvão" do século XXI. A abundância de dados que produzimos diariamente garante uma fonte quase inesgotável de informações para o treinamento de sistemas de IA. Entretanto, o uso de dados pessoais sem consentimento é uma das principais preocupações encontradas na literatura (Ekstrand et al., 2018).
Responsabilização	Como tornar a indústria da IA responsável por suas tecnologias. Por exemplo, no caso de veículos autônomos, que tipo de garantias e responsabilidades as empresas que desenvolvem veículos autônomos devem fornecer a seus clientes e à sociedade em geral (Maxmen, 2018)?
Confiabilidade	Confiabilidade é um princípio ético próximo da transparência. Este princípio defende a ideia de que sistemas de IA devem ser robustos. Dependendo do tipo de modelo, e contexto que tal modelo está inserido (e.g., automatizando tomadas de decisão do sistema judiciário), é de suma importância que tais sistemas sejam resilientes a, por exemplo, ataques adversariais (Krafft et al., 2020).
Beneficência/Não-maleficência	Este princípio defende que a inteligência artificial seja utilizada para promover o "Bem". Como o "Bem" é um conceito difícil de especificar, muitos consideram não-maleficência (e.g., IA não deve causar danos) como uma melhor especificação. Este princípio está muito próximo do que chamamos de Segurança da IA (Amodei et al., 2016).
Liberdade/Autonomia	Este princípio defende a ideia de que liberdade/autonomia (i.e., a experiência de que

	somos donos e responsáveis por nossas próprias escolhas e preferências) é fundamental para o bem-estar psicológico humano. Sistemas de IA não devem remover nossa autonomia, mas sim empoderá-la (Calvo et al., 2020).
Dignidade	Este princípio se refere ao valor inerente (e à vulnerabilidade inerente) do indivíduo humano. Algo que deve ser (a dignidade humana) inviolável. Sistemas de IA devem ser desenvolvidos de modo a promover um ecossistema que garanta que indivíduos sejam vistos, ouvidos, escutados, tratados com justiça, reconhecidos, compreendidos, e se sintam seguros (Ruster, 2021).
Sustentabilidade	Este princípio pode ser entendido como uma forma de "justiça intergeracional". Sustentabilidade descreve nossa obrigação ética para com as gerações futuras. Obrigação essa de garantir e preservar suas condições de vida, como, por exemplo, através do uso cuidadoso de nossos recursos naturais (Krafft et al., 2020).
*Solidariedade**	Este princípio pode ser entendido como o compartilhamento da prosperidade criada pela IA. Devemos implementar mecanismos para redistribuir o aumento da produtividade, compartilhar novos encargos e responsabilidade, e nos certificar de que a IA não irá aumentar a desigualdade de nosso mundo (Luengo-Oroz, 2019).
*Diversidade**	Este princípio pode ser entendido como a defesa e valorização das diferentes formas pelas quais a entidade humana pode vir a se expressar, por qualquer grupo ou identidade que deseje. Sistemas de IA devem ser desenvolvidos de forma a proteger e valorizar nossa diversidade (AIRES at PUCRS, 2021).
*Inclusão**	Sistemas de IA devem ser desenvolvidos de forma a "incluir", e não excluir. Este princípio defende o acolhimento de todas as formas pelas quais a entidade humana pode vir a se expressar, independentemente de filiações, grupos ou identidades específicas (AIRES at PUCRS, 2021).

* A respeito dos princípios de *Solidariedade*, *Diversidade* e *Inclusão*: esses são os princípios menos levantados pelo estado atual da Ética da IA. Contudo, achamos necessário apontá-los como importantes, e incluí-los dentro desta pequena e incompleta lista.

Ao mesmo tempo, em 2017, a Associação de Normas da IEEE publicou a segunda versão do documento *"Ethically Aligned Design: A Vision for*

Prioritizing Human Well-being With Artificial Intelligence and Autonomous Systems". Tal documento sugere diversas metodologias para orientar a pesquisa ética em projetos que busquem desenvolvimento de inteligência artificial, defendendo os valores humanos delineados pela Declaração Universal das Nações Unidas dos Direitos Humanos. O documento até mesmo orienta certas diretrizes e recomendações para o desenvolvimento de *"IAG eticamente alinhada"* (p. 73-82). [6]

Contudo, existem várias críticas levantadas contra este tipo de metodologia baseada em princípios abstratos (Principialismo), que sem uma tradução para a prática do desenvolvimento de sistema inteligentes, correm o risco de serem categorizadas como apenas um mero "teatro de ética", i.e., um discurso moral com pouca (ou nenhuma) intenção de resolver problemas do mundo real (Calo, 2017; Rességuier & Rodrigues; 2020; Corrêa & De Oliveira; 2021).

Nas palavras de Mittelstadt (2019, p. 503):

> Declarações baseadas em conceitos normativos vagos escondem pontos de conflito político e ético. A "justiça", a "dignidade" e outros conceitos abstratos são exemplos de "conceitos essencialmente contestados" que têm muitos possíveis significados conflitantes que requerem interpretação contextual [...] Na melhor das hipóteses, esta ambiguidade conceitual permite uma especificação sensível ao contexto das exigências éticas para a IA. Na pior das hipóteses, ela mascara desacordos fundamentais de princípios, e conduz a ética da IA em direção ao relativismo moral. No mínimo, qualquer compromisso alcançado até agora em torno de princípios fundamentais para a ética da IA não reflete um consenso significativo

[6] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. Disponível em: http://standards. ieee.org/develop/indconn/ec/autonomous_ systems.html.

> sobre uma direção prática comum para o “bom” desenvolvimento e governança da IA.

Assim, é importante ressaltar que ainda existem poucas propostas de como devemos implementar princípios éticos e diretrizes normativas na prática do desenvolvimento de sistemas de IA. O objetivo deste trabalho é tentar diminuir esta lacuna entre o discurso e a práxis. Entre princípios abstratos e implementação técnica.

Para darmos início a este guia, na próxima seção, iremos investigar como é o ciclo de desenvolvimento de modelos treinados por aprendizagem de máquina, a fim de melhor entendermos quais os pontos onde problemas podem surgir.

The AI Robotics Ethics Society®

O “Ciclo de Vida” de um Sistema de Aprendizagem de Máquina

O “ciclo de vida” de um modelo treinado por aprendizagem de máquina, i.e., a concepção, criação, implantação e monitoramento de tal sistema, não pode ser isolado de interferências humanas, já que tais sistemas são concebidos, implantados, utilizados e monitorados por humanos. Ananny & Crawford, (2018) sugerem que sistemas algorítmicos são conjuntos de atores humanos e não-humanos, emaranhados em uma dinâmica que gera efeitos não-determinísticos, e em nossa opinião, esta é uma excelente definição.

Assim, para compreender e prospectar as implicações éticas de tais sistemas, é necessário compreender como todo o sistema funciona, i.e., o conjunto total de atores humanos e não-humanos que o compõem.

Sistema criados por aprendizagem de máquina (especialmente aprendizagem supervisionada e suas variantes) geralmente seguem o seguinte ciclo:

- *Coleta de dados e pré-processamento:* antes que qualquer análise ou aprendizagem possa acontecer, precisamos de dados. O conjunto de dados para treinamento é geralmente criado a partir de duas hipóteses: (1) *você supõe que suas saídas podem ser previstas dadas suas entradas*; (2) *você supõe que os dados disponíveis são suficientemente informativos para aprender a relação entre entradas e saídas*. Algumas vezes podemos aceitar tais premissas (e.g., tamanho de sapato está geralmente relacionado com a altura do indivíduo), e às vezes não podemos (e.g., o histórico de valor de algum ativo, como ações da Apple, pode

não estar correlacionado com seu valor futuro). Se você quiser criar um sistema para prever a chance de um paciente desenvolver diabetes, você poderia criar um banco de dados com base nos usuários da rede pública de saúde que possuem diabetes tipo 2. Você poderia utilizar certas características (features) que você acredita que estão correlacionadas com diabetes tipo 2 (e.g., peso, idade, IMC, histórico familiar, prática de atividade física), junto com amostras de pessoas que possuem diabetes tipo 2 (dados rotulados) para criar um modelo que, dado os valores de entrada que você estabeleceu, prevê qual a chance desse indivíduo desenvolver, ou não, diabetes tipo 2. Quase que 80% de todo o trabalho de um engenheiro de aprendizagem de máquina está em criar um bom conjunto de dados;

- *Desenvolvimento do Modelo:* após definirmos nosso conjunto de dados, o dividimos em três grupos: treinamento, validação e testagem. É considerada uma "boa-conduta" (se não senso-comum) não misturar o conjunto de dados de treinamento com os que serão utilizados para validação e testagem. O que queremos é um modelo que generalize suas previsões para amostras novas, e não um modelo que simplesmente decore os dados apresentados (i.e., sobreajuste). Após essa divisão, definimos a arquitetura de nosso modelo (e.g., feedforward neural network), nossa função objetiva (e.g., prever os indivíduos mais suscetíveis a diabetes tipo 2), nossa função de perda (e.g., entropia cruzada binária), otimizador (e.g., gradiente descendente estocástico) e métrica de avaliação (e.g., acurácia, precisão, recall, AUC). Após, iremos iterativamente ajustar os parâmetros de nosso modelo (e.g., número de nodos na camada oculta, taxa de aprendizagem do otimizador, uma função de perda diferente) de modo a melhorar o resultado de nossa métrica de avaliação nos dados de validação. Nesta etapa, (re)treinamos nosso modelo até que estejamos satisfeitos com seu resultado em validação;

- *Avaliação do Modelo:* nós treinamos o modelo com a porção dos dados dedicada ao treinamento, e avaliamos sua performance para ajuste dos parâmetros com a porção dedicada a validação. Em seguida, quando já nos encontramos satisfeitos com o modelo final, o avaliamos com a porção dos dados dedicada à fase de teste. Esse é o momento que avaliamos o poder preditivo de nosso modelo para com dados nunca vistos antes. Para isso, também é importante decidir que tipo de métrica de performance escolheremos para avaliar nosso modelo. Também podemos avaliar nosso modelo em testes-padrão (benchmarks), se estes existirem;
- *Pós-processamento:* nesta fase, precisamos adaptar a saída de nosso modelo para o problema que estamos lidando. Por exemplo, se definirmos que a saída de nosso modelo (i.e., o modelo que infere a chance de um indivíduo desenvolver diabetes tipo 2) é uma medida de probabilidade entre 0 e 1, mas queremos uma resposta categórica (i.e., "Sim", "Não", "Inconclusivo"), precisamos estimar um limiar de decisão (e.g., mais de 80% de confiança é igual a uma classificação conclusiva);
- *Implantação do modelo:* nesta etapa, muitos ajustes devem ser feitos para tornar o modelo "ergonômico", facilitando sua interação para com seus usuários. Por exemplo, ferramentas de transparência podem ser implementadas no modelo, em casos onde a interpretabilidade seja vital (e.g., o sistema VICTOR, utilizado pelo Supremo Tribunal Federal). Implantação é sempre um momento sensível do ciclo de vida desses sistemas. Já que o ambiente de treinamento (geralmente) não é uma representação fiel do ambiente de implantação, comportamentos inesperados/indesejados podem ocorrer apenas nessa fase (e.g., um sistema de recomendação de roupas que foi treinado

ingenuamente no verão, e é incapaz de recomendar peças de roupa "úteis" durante o inverno);

- *Monitoramento:* após a implantação do modelo, é necessário monitorar seu comportamento, de modo a assegurar que o sistema cumpre a função para a qual foi desenvolvido, e não resulta em nenhum tipo de comportamento que possamos vir a considerar como indesejado/inseguro (e.g., um modelo de predição de doenças que possui uma baixa performance para com um grupo específico deve, em tese, ser retirado de circulação e aperfeiçoado).

Com o ciclo de desenvolvimento, implantação e monitoramento desse tipo de sistema definido, investigaremos na próxima seção "onde" problemas podem surgir e comprometer um modelo criado por aprendizagem de máquina.

Fontes de Problemas

Como podemos ver na seção anterior, muitas pressuposições e escolhas serão tomadas antes de implantarmos nosso modelo para agir no mundo real. Desenvolvedores, engenheiros de aprendizagem de máquina, cientistas de dados, todos estes atores estarão ativamente influenciando a forma que o modelo irá tomar durante seu desenvolvimento, seja na construção dos conjuntos de dados (treinamento, validação e testagem), escolha e confecção das características (feature engineering), definição dos parâmetros do modelo, escolha do método de avaliação, etc. Durante esse longo processo, muitas "más" decisões podem influenciar negativamente o modelo final.

Como mencionado na seção inicial, a origem desses problemas não é simplesmente "dados enviesados". Em primeiro lugar, bancos de dados não são estruturas estáticas, divorciadas de intenções e contextos sociais/históricos do qual surgiram. Mascarar os efeitos colaterais gerados por tais sistemas como apenas "dados enviesados" é ofuscar a complexidade de como tais sistemas podem vir a ser comprometidos ao longo de seu ciclo de vida. Ao mesmo tempo, é ofuscar a nossa parcela de responsabilidade para com o problema.

Suresh & Guttag (2021), em seu estudo "*Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle*", fornecem uma estrutura de análise que identifica sete fontes distintas de danos que podem comprometer o comportamento desse tipo de sistemas, desde a coleta de dados até sua implantação:

- *Vieses históricos:* este tipo de problema ocorre pelo fato de que o nosso próprio mundo, como é ou foi, é falho. Assim, mesmo que o modelo seja uma representação perfeita do ambiente, ele ainda

poderá gerar danos, pois representa um ambiente imperfeito. Por exemplo, Brown et al. (2020, p. 36-37) reportam que seu modelo de linguagem (GPT-3) associa adjetivos pejorativos, sexistas e misóginos com mais frequência há mulheres do que homens (i.e., uma reflexão dos textos, e cultura, que encontramos pela internet);

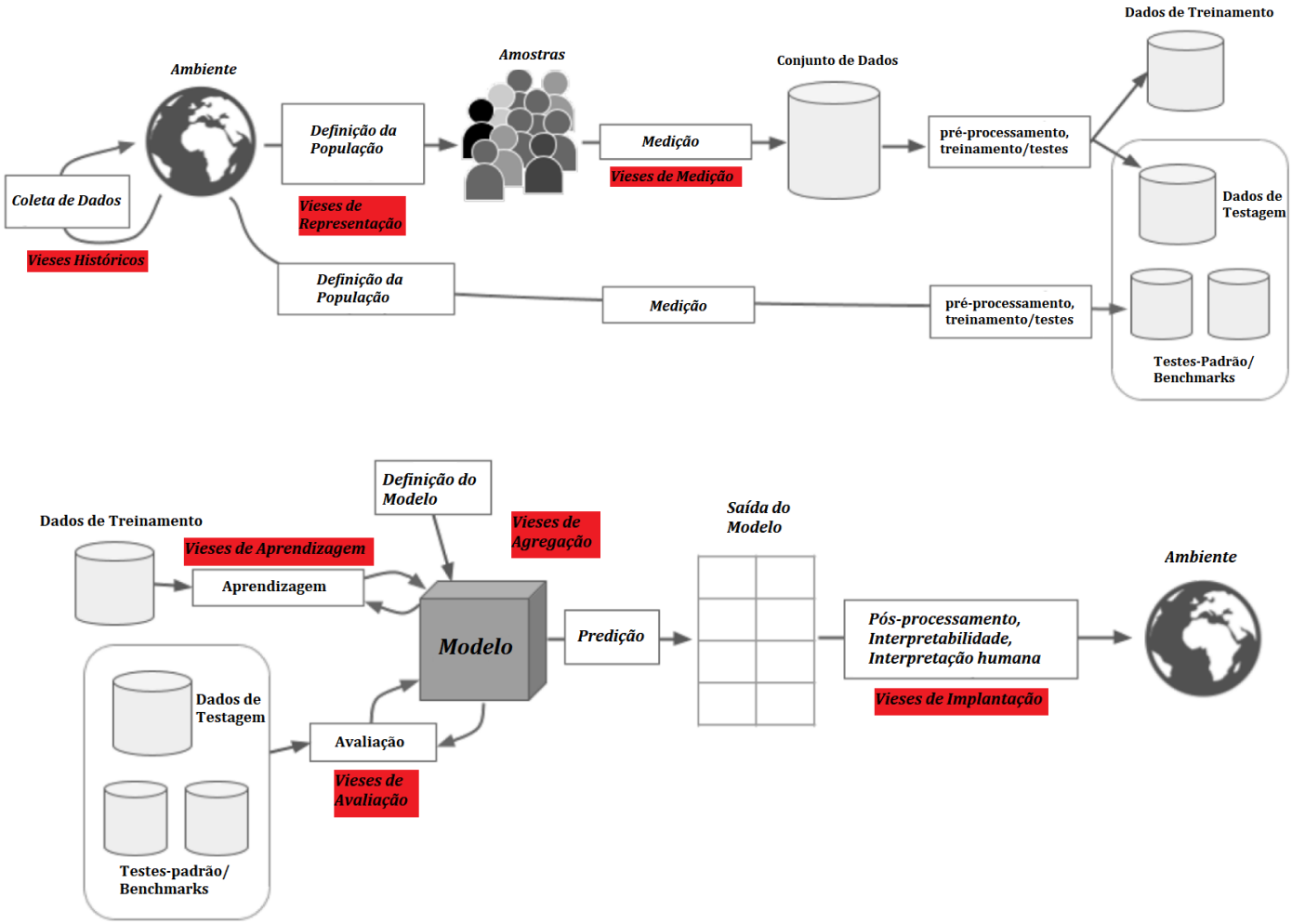
- *Vieses de representação:* este problema ocorre quando os dados utilizados para o treinamento do modelo não representam a população ou ambiente em que o modelo irá agir. Quando treinando um modelo para prever o desenvolvimento de diabetes tipo 2, talvez não haja dados o suficiente para representar todos os possíveis grupos de interesse (homens, mulheres, idosos, crianças, etc.). Ou, um software de reconhecimento de imagem utilizado por um carro autônomo treinado em ambientes urbanos, pode operar de forma falha quando operando em regiões rurais;
- *Vieses de medição:* quando escolhemos características (features) para serem utilizadas por algum resultado (e.g., IMC, peso, altura, histórico familiar), estamos supondo que tal característica/quantidade é representativa daquilo que queremos prever ou classificar (diabetes tipo 2). Mas esse nem sempre é o caso, pois tais "proxies" podem ser apenas aproximações de uma realidade mais complexa. Por exemplo, se nosso modelo atribui muito peso a variável "IMC" para a tarefa de predição de diabetes tipo 2, sujeitos com um grande volume muscular (e.g., fisiculturistas) poderão ser classificados falsamente como potenciais desenvolvedores de diabetes tipo 2. "QI" (i.e., quociente de inteligência) pode não ser um bom parâmetro para se avaliar sucesso acadêmico, que muitas vezes depende de outros fatores difíceis de serem mensurados (e.g., motivação, capacidade de se relacionar, habilidades organizacionais);
- *Vieses de agregação:* este tipo de problema ocorre quando grupos diferentes são unidos em um único conjunto de dados. Contudo, o modelo treinado não atua eficientemente com algum (ou nenhum) dos grupos. Por exemplo, sabe-se que homens têm duas vezes mais

chances de ter um ataque cardíaco ao longo da vida. Um modelo treinado com um conjunto de dados misto (i.e., sem diferenciar o sexo das amostras) para prever a chance de um paciente ter um ataque cardíaco, pode vir a ser ineficiente para com algum dos sexos (ou ambos). Idealmente, modelos devem ser treinados para atender grupos específicos (quando necessário);

- *Vieses de aprendizagem:* a escolha da função de perda (e.g., erro médio quadrático, entropia cruzada binária, entropia cruzada categórica) e métrica de performance (e.g., acurácia, precisão, recall, AUC) pode influenciar o tipo de saída que nosso modelo gera, e como interpretamos sua performance. Por exemplo, se temos uma aplicação para qual classificações negativas falsas podem gerar um "grande custo" (e.g., falso negativo para HIV), talvez não devamos utilizar acurácia para medir sua performance, mas sim recall;
- *Vieses de avaliação:* nem sempre os dados utilizados na fase de testagem (ou teste-padrão/benchmark) representam uma boa métrica de avaliação para o domínio em que o modelo será implantado. Por exemplo, você pode ter desenvolvido um modelo de reconhecimento facial com uma excelente performance em sua fase de teste. Contudo, o benchmark que você utilizou tem uma baixa representação da população parda (e.g., 4%), e o modelo irá atuar em um domínio onde grande parte da população é parda (e.g., Brasil);
- *Vieses de implantação:* este tipo de problema ocorre quando o modelo é utilizado de forma diferente, ou além, daquilo que foi originalmente desenvolvido para fazer. Muitos dos modelos criados por aprendizagem de máquina não são "totalmente autônomos", mas se encontram como parte de um processo sociotécnico onde intenções e desejos humanos fazem parte. Por exemplo, sistemas

de avaliação de risco são utilizados no sistema penal americano para prever a probabilidade de uma pessoa cometer um crime futuro (i.e., reincidência criminal). Contudo, uma instanciação perversa desta ferramenta seria a utilizar para determinar a *duração de uma sentença com base no risco provável de reincidência* (Collins, 2018).

Abaixo temos um diagrama descrevendo o ciclo de vida de um modelo desenvolvido por aprendizagem de máquina, e onde os vieses descritos acima se manifestam.



Fontes de problemas durante o desenvolvimento e implantação de um modelo (Suresh & Guttag, 2021).

É importante ressaltar que dependendo da aplicação de um modelo, os tipos de problemas citados acima não irão se manifestar de nenhuma

forma prejudicial, tornando uma "análise ética" desnecessária. Por exemplo, um modelo criado para otimizar processos industriais (e.g., controle de qualidade, controle de estoque), que não afetam a vida das pessoas de uma maneira significativa, não necessita do mesmo nível de análise que modelos que interagem diretamente com pessoas, necessitam. Em outras palavras, não existe, por exemplo, "quebra de privacidade" em situações onde um modelo deve classificar uma fruta como "própria para consumo" ou "imprópria para consumo".

Contudo, se durante uma inspeção inicial for revelado que há questões éticas a se considerar, a organização e desenvolvedores responsáveis devem realizar uma avaliação ética completa do modelo.

Em suma, *o contexto é o que definirá a norma.* Não existe uma "Única e Verdadeira Teoria Moral" que se aplique para toda e qualquer aplicação de um modelo algorítmico.

Na próxima seção, agora que já sabemos diversos tipos de problemas que podem interferir com o desenvolvimento e implantação de um modelo criado por aprendizagem de máquina, iremos explorar algumas das definições de "justiça" (às vezes também referida como "igualdade" ou "fairness") algorítmica.

Definindo “Justiça” em Aprendizagem de Máquina

Levando todos os exemplos de vieses citados na última seção, o que gostaríamos de fazer, idealmente, é desenvolver um modelo “justo”, i.e., um modelo que realiza sua função livre de discriminações e preconceitos. Há um corpo crescente de trabalho sobre “algoritmos justos” sendo publicado, e podemos definir “algoritmos justos”, no contexto da aprendizagem de máquina, como um modelo que satisfaça alguma noção particular de “justiça”.

Contudo, dependendo de como formalizamos “justiça” ou “equidade”, diferentes decisões/classificações/predições serão definidas como “justas”. Decisões essas que podem conflitar com outras formalizações particulares de “o que queremos dizer com justo”:

- Justiça significa alcançar paridade entre os grupos demográficos de uma população;
- Justiça significa satisfazer as preferências dos grupos demográficos de uma população;
- Justiça significa beneficiar igualmente todos os grupos demográficos de uma população;
- Justiça significa impactar (i.e., oposto de beneficiar) igualmente todos os grupos demográficos de uma população;
- Justiça significa julgar por trás do véu da ignorância;

Qual seria a melhor definição para aplicarmos no contexto da aprendizagem de máquina? Em primeiro lugar, precisamos definir justiça, em suas mais variadas formas, em termos estatísticos, i.e., como as inferências estatísticas de um modelo podem ser realizadas de modo a respeitar noções específicas de justiça. Vejamos algumas dessas possíveis definições:

- *Véu da Ignorância:* um modelo satisfaz essa condição se todos os atributos sensíveis (i.e., atributos para o qual a não-discriminação deve ser estabelecida) de suas amostras não são explicitados ao modelo, i.e., o modelo não possui acesso a informações como raça, etnia, cor, nacionalidade, sexo, orientação sexual, etc.

Esta abordagem pode ser remetida à definição de Justiça defendida por John Rawls (1999), em sua obra seminal "A Theory of Justice". Um dos problemas com essa abordagem é que precisamos definir quais proxies podem ser utilizados por um modelo para identificar (e discriminar) amostras. Mesmo que atributos sensíveis sejam velados do modelo, um preditor ainda poderia inferir e discriminar populações marginalizadas com base em informações não-sensíveis. Por exemplo, se um banco utiliza um modelo para auxiliar na avaliação de conceção de linhas de crédito, e a cidade/região possui um certo nível de segregação racial em sua distribuição de moradores ("bairros nobres" versus "favelas"), atributos não-sensíveis, como CEP, podem ser utilizados para discriminar indivíduos que morem em certas localizações.

Além disso, certos estudos apontam que o véu da ignorância pode vir a ser mais discriminatório do que "justiça consciente" (i.e., quando levamos em conta atributos sensíveis em um julgamento) (Sen, 1990; Bonilla-Silva, 2003; Fryer et al., 2008). Ao mesmo tempo, essa abordagem parece ir contra princípios de "reparação e auxílio" de populações historicamente marginalizadas.

- *Justiça por Conhecimento:* um modelo satisfaz essa condição se o modelo produz a mesma saída para indivíduos semelhantes. Ou seja, se duas amostras possuem um número mínimo de características similares, ambas as amostras serão classificadas de forma igual.

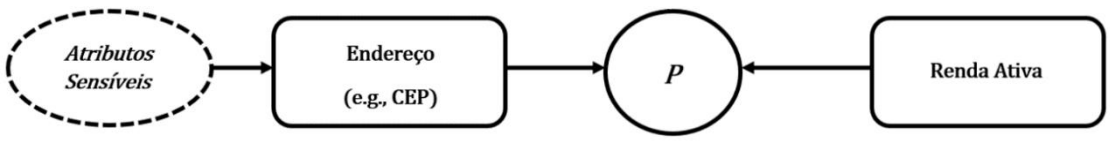
Essa definição é uma variação mais elaborada do critério anterior ("Véu da Ignorância"), onde definimos como justo um modelo que trata indivíduos semelhantes de forma semelhante. Para colocarmos em prática esta definição, primeiro precisamos: (1) definir uma métrica de distância para medirmos a similaridade entre duas amostras; e (2) definir qual a distância mínima para que duas amostras sejam classificadas da mesma forma. Por exemplo, uma possível métrica de distância f poderia definir que a distância entre dois sujeitos, i e j, como 0 se todos os atributos não-sensíveis são idênticos e 1 se algum atributo não-sensível for diferente.

Contudo, não é nada claro como criar uma função que toma como argumento as características de duas amostras e calcular sua distância sem gerar classificações discriminatórias. Por essa definição, nós apenas "passamos o problema" de se fazer uma classificação justa do modelo para a métrica de distância.

- *Igualdade Contrafactual:* um modelo satisfaz a condição de igualdade contrafactual se sua classificação não é alterada, mesmo que os atributos sensíveis da amostra *fossem diferentes*.

Isso é a mesma coisa que dizer que o modelo é "justo" se, por exemplo, autoriza a abertura para uma linha de crédito para o sujeito A, que possui o atributo sensível i, mesmo que o sujeito A possuísse o atributo sensível j.

Para implementar um modelo que satisfaça essa condição, primeiro precisamos determinar o que seria uma "influência contrafactual" (i.e., como atributos sensíveis podem influenciar atributos não-sensíveis a determinar o resultado de um classificador). Uma forma de fazermos esse tipo de análise é utilizando gráficos/diagramas causais (Pearl, 1995; Kilbertus et al., 2017).



No gráfico acima, atributos sensíveis (e.g., raça) poderiam ser inferidos por "Endereço", que agiria como um proxy para atributos sensíveis em uma cidade onde ocorre segregação racial na distribuição de seus moradores. Assim, podemos definir que um modelo causal é justo por igualdade contrafactual se nenhum atributo sensível pode influenciar um proxy que afeta diretamente a decisão final (P) do modelo (o modelo acima não satisfaz essa condição).

A fim de satisfazer essa condição de justiça, seria necessário não utilizar (como características de classificação) todos os nodos descendentes de nodos contendo atributos sensíveis. Contudo, em aplicações reais, a grande maioria dos atributos sensíveis são nodos cujos descendentes abrangem praticamente todo o gráfico causal.

- *Paridade estatística/demográfica:* um modelo satisfaz a condição de paridade estatística/demográfica se modelo é capaz de gerar resultados iguais, ou quase iguais, para membros de grupos com atributos sensíveis diferentes.

Paridade estatística para grupos pode ser remetida a uma noção de justiça distributiva igualitária-coletivista (Hofstede et al., 2010; Hirose, 2014). Uma das críticas levantadas contra essa formalização é que ao impormos paridade estatística ao nosso modelo, poderemos gerar: (1) um modelo que produz resultados incorretos (i.e., o modelo irá priorizar paridade estatística em detrimento de seu desempenho); ou (2) o modelo pode vir a discriminar contra indivíduos mais "qualificados" (Luong et al., 2011, Dwork et al., 2012). Por exemplo, um modelo designado a auxiliar no processo de seleção para uma vaga de emprego poderia vir a priorizar paridade estatística sobre outros atributos relevantes (e.g., "gênero sobre currículo").

Assim, é importante que entendamos que a imposição de métricas de justiça (e.g., paridade estatística/demográfica) deve ser vista como um

compromisso entre “desempenho” e “justiça”, já que estaremos intencionalmente introduzindo imparcialidades que desviam da distribuição de dados original.

Para certas aplicações, isto não seria desejável. Por exemplo, um classificador de sentimento (uma tarefa de aprendizagem de máquina comum em NLP) enviesado para classificar palavras como “suicídio”, “depressão”, “solidão”, como um texto contendo sentimento negativo é desejável. Contudo, modelos enviesados a favor, ou detrimento, de classes sociais, raças ou gêneros, não são desejáveis.

Podemos também definir paridade estatística em termos de resultado previsto:

- *Paridade Preditiva:* Um modelo satisfaz a condição de paridade preditiva se a precisão do modelo é igual entre diferentes grupos. Ou seja, se o modelo determina com 90% de precisão[7] que um indivíduo é um “bom candidato para um empréstimo” (i.e., o empréstimo não será prejudicial para o banco ou para o sujeito), essa medida de precisão deve ser independente do valor de atributos sensíveis.

Uma das dificuldades em se implementar um modelo que satisfaça a condição de paridade preditiva entre diferentes grupos é que nem sempre grupos são igualmente representados em conjuntos de dados. Por exemplo, dos grandes conjuntos públicos de imagens de rosto (e.g., UTKFace, CelebA, LFWA+), existe uma forte tendência em prol de rostos caucasianos, enquanto que outras raças (e.g., indígenas) são significativamente sub-representadas (Kärkkäinen & Joo, 2019). Para que um modelo alcance uma performance equilibrada entre grupos, é importante que haja exemplos suficientes para que o sistema possa aprender um bom modelo.

[7] Precisão = Verdadeiros Positivos / Verdadeiros Positivos + Falsos Positivos.

- *Probabilidades Equalizadas:* esta condição de justiça algorítmica pode ser interpretada como uma extensão da condição de paridade preditiva. Um modelo que satisfaz essa condição é um modelo que tem uma taxa de verdadeiros e falsos positivos igual, independentemente do valor de atributos sensíveis.

Isso significa que a chance de um indivíduo com uma boa pontuação de crédito receber uma classificação positiva (i.e., ter uma nova linha de crédito aprovada), e a chance de um indivíduo com uma pontuação ruim de crédito receber uma classificação positiva, é igual e independente do grupo (i.e., atributos sensíveis) que esse indivíduo pertence. Em outras palavras, ambos os membros de *i* e *j* têm a mesma chance de receberem uma classificação positiva (seja ela correta ou não).

Certamente que essas não são as únicas definições existentes de justiça algorítmica, e outras definições podem ser encontradas na literatura sobre "machine learning fairness" (Chouldechova, 2016; Hardt et al., 2016, Corbett-Davies et al., 2017; Galhotra et al., 2017; Kilbertus et al., 2017; Verma & Rubin, 2018, Gajane & Pechenizkiy, 2018, Mehrabi et al., 2019). Contudo, o ponto que pretendemos deixar claro é:

- *Não existe uma única definição de o que é justo.*

Organizações preocupadas em desenvolver sistemas de IA "justos" (i.e., capazes de mitigar o surgimento de certos vieses ao longo do ciclo de vida do modelo criado) devem primeiro estabelecer qual a melhor definição de "justo" que se aplica ao problema que seu modelo irá enfrentar. Certas definições produziram melhores resultados para determinados tipos de problemas. Para certas aplicações, talvez seja importante "obscurecer" todas as formas de atributos sensíveis. Para outras, talvez seja melhor priorizar a paridade estatística em detrimento da eficiência de classificação.

Em suma, *a correta definição de "Justiça" depende do contexto.* Contudo, não seria possível aplicarmos todas as definições sugeridas como restrições de equidade para o mesmo modelo? Infelizmente, existem limitações de como podemos restringir as predições de modelos de inferência estatística. Vejamos algumas dessas restrições na próxima seção.

Resultados de Impossibilidade em Ética da IA

Muitas das definições apresentadas na última seção podem parecer similares ou variantes de um de um mesmo objetivo geral, i.e., que as inferências de um modelo probabilístico sejam independentes de certos atributos sensíveis, como gênero, raça, orientação sexual, etc. Contudo, quando tentamos calibrar nosso modelo de modo que ele satisfaça múltiplas noções de igualdade e paridade estatística, chegamos a certos resultados de impossibilidade.

Desde 2016, graças ao trabalho de Kleinberg et al. (2016), já sabemos que certas noções de justiça, no contexto de classificações probabilísticas, são incompatíveis umas com as outras, i.e., não podemos satisfazer todas ao mesmo tempo. Existem certas arbitragens inevitáveis entre diferentes definições, independentemente do contexto específico e do método utilizado para se chegar a uma classificação probabilística.

Em aprendizagem de máquina, quando projetamos um modelo para fins de classificação/predição, utilizamos um grupo de dados (x_i) rotulados (y_i) para treinarmos nosso modelo. O objetivo que nosso modelo deve cumprir é encontrar uma função $f: X \to Y$ que aproxime a verdadeira distribuição conjunta de amostras e rótulos ($X \times Y$). Assim, o objetivo do modelo pode ser definido em termos de minimização de risco empírico, i.e., a diminuição da lacuna entre as predições do modelo e os verdadeiros rótulos de suas amostras.

Modelagem estatística para fins de minimização de risco empírico pode ser considerada como uma condição ortogonal a qualquer noção de justiça que faça o modelo desviar da verdadeira distribuição conjunta de

amostras e rótulos. Em outras palavras, técnicas padrão de minimização de perda (e.g., entropia cruzada binária) e otimização (e.g., gradiente estocástico descendente) buscam minimizar o risco empírico, e não aderir a noções particulares de justiça. Assim, um classificador justo seria aquele que diminui a lacuna entre "predições justas" e "predições empíricas" (Saravanakumar, 2021).

Ao projetarmos um classificador justo, o problema que queremos evitar é que atributos sensíveis interfiram na classificação do nosso modelo. Podemos dizer que um atributo sensível (a) é uma possível fonte de viés, apenas se tal atributo está estatisticamente correlacionado com a predição ($\hat{y}$) de nosso modelo($P_a[\hat{y}] = P(a)$). Caso contrário, o atributo sensível não irá interferir na inferência do modelo em questão, pois ele não está correlacionado com a predição, ou valor verdadeiro, da amostra sendo considerada.

Pelo que podemos ver na última seção, definir um algoritmo de inferência estatística "justo" é equivalente a definir algum critério de calibração que se enquadre em alguma das diversas definições de "justiça", "equidade" e "igualdade" encontradas na literatura. Os resultados de impossibilidade de Kleinberg et al. (2016) se aplicam a três dessas definições, ditando que, salvo casos ideais, nenhum modelo de inferência estatística pode satisfazer os três seguintes critérios de calibração:

A. *Calibração dentro de grupos:* um modelo satisfaz essa condição se para cada possível grupo (e.g., i e j), o modelo classifica os membros de i e j que satisfazem a condição positiva para uma determinada classe com a mesma chance, i.e., tanto os indivíduos do grupo i quanto do grupo j que possuem a mesma probabilidade de serem classificados positivamente, terão a mesma chance de serem classificados positivamente pelo modelo.

B. *Equilíbrio para a classe positiva:* um modelo satisfaz essa condição se a chance de um indivíduo ser classificado para a classe positiva é independente de seu grupo. Um modelo que não satisfaz essa condição é um modelo que privilegia (i.e., classifica positivamente

com maior chance) membros de um grupo (i) em detrimento do outro (j).

C. *Equilíbrio para a classe negativa:* é a condição inversa da definição anterior. Um modelo satisfaz essa condição se a chance de um indivíduo ser classificado como uma classe negativa é independente de seu grupo. Corolário, um modelo que não satisfaz essa condição é um modelo que privilegia membros de um grupo em detrimento do outro.

O critério A pode ser definido como uma forma mais restrita da condição de Paridade estatística/demográfica. Já os critérios B e C podem ser interpretados como versões das condições de Paridade Preditiva e Probabilidades Equalizadas. Essas três definições de justiça algorítmica são algumas das mais aceitas e estudadas pela comunidade, também sendo essas as vítimas deste resultado de impossibilidade.

De acordo com os teoremas de impossibilidade de Kleinberg et al. (2016), só existem duas exceções a essa regra:

- *Casos Ideais:* os únicos exemplos de problemas nos quais existe uma classificação probabilística que satisfaz as condições de justiça A, B e C, são quando: (1) o modelo de inferência é perfeito (i.e., a distribuição conjunta de $X \times Y$ é perfeitamente conhecida, e $P[\hat{y}]$ é igual a 0 ou 1 para todos x_i); ou (2) o modelo de inferência possui taxas de base iguais (i.e., a chance de uma amostra ser classificada como pertencente a classe positiva e negativa é igual).[8]

[8] Kleinberg et al. (2016) também provaram que seus resultados de impossibilidade podem ser estendidos para situações onde apenas aproximamos os casos ideais, i.e., o modelo apenas aproxima uma predição perfeita com um erro $\varepsilon > 0$, ou o modelo apenas aproxima uma proporção igual entre o total equilíbrio entre classes negativas e positivas, para qualquer $\delta > 0$.

Infelizmente, ambos os casos ideais não são a “norma” em termos de modelos probabilísticos criados por aprendizagem de máquina. Se soubéssemos a distribuição perfeita de todas as possíveis amostras, nós não precisaríamos de aprendizagem de máquina, pois temos acesso a um oráculo. E um modelo com uma taxa de equilíbrio entre classes negativas e positivas igual, geralmente, não irá representar a verdadeira distribuição de dados. Diversas situações e aplicações não podem ser decididas pelo jogar de uma moeda justa (mesmo que isso seja, “estatisticamente”, o mais justo a se fazer).

Utilizemos novamente nosso exemplo de um indivíduo que procura um banco para tentar abrir uma nova linha de crédito, e uma de suas avaliações será realizada por um modelo de inferência estatística (e.g., um modelo de aprendizagem de máquina treinado de forma supervisionada), que resultará em sua “pontuação de crédito”.

Provavelmente, a distribuição de amostras não é uniforme, ou seja, o ambiente não é formado de 50% de pessoas com uma boa pontuação de crédito (i.e., uma nova linha de crédito seria benéfica para o banco e para o indivíduo) e 50% de pessoas com uma má pontuação de crédito (i.e., uma nova linha de crédito seria danosa tanto para o banco quanto para o indivíduo). Da mesma forma, a distribuição de atributos sensíveis entre as duas classes (por simplicidade, estamos imaginando um problema de classificação binária) provavelmente não é uniforme (i.e., talvez a distribuição real de “indivíduos com uma boa pontuação de crédito” favoreça mulheres).

Assim, o que os resultados de impossibilidade nos indicam é que com exceção dos dois tipos de casos ideais, pelo menos duas das seguintes propriedades indesejáveis devem se manter, pois nenhum modelo de inferência probabilística pode satisfazer, simultaneamente, os critérios de calibração A, B e C, i.e., somos apenas capazes de satisfazer um critério à custa de dois:

1) *Violação de Paridade estatística/demográfica:* os resultados do classificador/preditor/modelo são sistematicamente enviesados, para cima ou para baixo, para pelo menos um grupo;
2) *Violação de Paridade Preditiva:* a taxa de classificações para a classe positiva é sistematicamente enviesada, atribuindo maior probabilidade para a classe positiva para pelo menos um grupo;
3) *Violação de Probabilidades Equalizadas:* a taxa de classificações para a classe negativa é sistematicamente enviesada, atribuindo maior probabilidade para classe negativa para pelo menos um grupo.

A troca entre essas três condições não é necessariamente um problema da aprendizagem de máquina, mas um fato sobre problemas de classificação probabilística que buscam modelar dados produzidos por fenômenos do mundo real. Essa impossibilidade não deve ser atribuída à falta de capacidade do modelo, mas sim às restrições do regime de geração de dados e as condições de igualdade e justiça que estipulamos. Outra forma de interpretar esse resultado é que o problema de justiça algorítmica não é exatamente um problema estatístico, mas um problema sociológico, já que as discrepâncias e vieses embutidos nos dados são apenas reflexos de uma sociedade desigual e imperfeita.

Assim, um engenheiro de aprendizagem de máquinas que constrói modelos para aplicações com possíveis impactos sociais deverá estar preparado para lidar com este fenômeno. Escolher uma métrica de justiça é também escolher quais as violações que estamos dispostos a fazer.

Não existe uma solução geral para esse problema. É da responsabilidade dos desenvolvedores e supervisores de um projeto que visa criar tais modelos, definir por qual régua normatizar seu sistema. Contudo, o que deve ser feito é: (1) investigar as limitações e possíveis vieses do modelo

gerado; (2) a disponibilização de tais informações (de forma transparente) a aqueles que irão utilizar (ser impactados por) tais tecnologias.

Um gerente de banco assistido por um modelo de IA deve saber, caso esse seja o caso, que seu modelo possui certos vieses em sua tomada de decisão. Tais vieses, e as medidas e escolhas que foram feitas para mitigar seus possíveis efeitos colaterais, devem estar explicitamente disponibilizados para o operador. Por exemplo, talvez o modelo possa vir com uma "bula" ou "carta", explicando os possíveis vieses que o modelo pode apresentar. Quando uma classificação para a amostra X for feita, talvez o modelo possa resultar em além de uma classificação, um aviso ("Atenção! Este modelo tende a gerar um percentual de Falsos Negativos sistematicamente enviesado para amostras com o seguinte atributo sensível: 'Divorciado(a)'.").

Na próxima seção, começaremos a apresentar algumas possíveis soluções (ferramentas e metodologias) para mitigar os problemas apresentados até agora.

The AI Robotics Ethics Society®

O papel do Engenheiro de Segurança da IA

Imagine que você é o responsável por uma divisão de ética e segurança da IA de uma empresa que produz soluções e produtos através de técnicas de aprendizagem de máquina. Seu dever é: (1) garantir que os modelos gerados pela sua empresa sigam certos protocolos de segurança; (2) garantir que possíveis efeitos colaterais sejam detectados e previstos antes que o modelo seja implantado para agir no ambiente; e (3) monitorar o comportamento do modelo "in the wild". Os problemas citados na última seção são algumas das preocupações que devem estar no seu radar:

- Como estão estruturadas em nossa fábrica social e política as diversas formas de opressão e preconceitos históricos característicos do contexto onde o modelo será implantado (i.e., vieses históricos)?
- Será que os dados utilizados para treinamento são representações fidedignas da população ou domínio de interesse? Haveriam grupos importantes, porém marginalizados, que não estão presentes nesse conjunto de dados (i.e., vieses de representação)?
- Será que os rótulos e características escolhidas são boas aproximações (proxies) daquilo que temos interesse de medir/classificar/prever (i.e., vieses de medição)?
- Dado um determinado problema, seria correto agregar grupos diferentes? Ou precisamos tratar cada grupo com respeito às suas especificidades (i.e., vieses de agregação)?

- Como o modelo pode ser utilizado para fins diferentes daqueles que foram definidos pelos desenvolvedores (i.e., vieses de implantação)? Quais tipos de ataques adversariais o modelo é mais suscetível?
- Quais métricas de justiça estão sendo seguidas? Quais condições de justiça algorítmica o modelo viola (i.e., Paridade estatística/demográfica, Paridade Preditiva, Probabilidades equalizadas)? Seriam restrições de equidade algorítmica algo necessário para a aplicação em questão?

Compreender onde a intervenção é necessária e como ela é viável pode informar as discussões sobre como danos podem ser mitigados versus quando é melhor não implantar um sistema de forma alguma. Comecemos a explorar algumas ferramentas qualitativas para auxiliar desenvolvedores a realizar tal análise.

The AI Robotics Ethics Society®

Ferramentas Translacionais

Ferramentas translacionais, no contexto da Ética da IA, são metodologias para auxiliar desenvolvedores a "traduzir" princípios éticos abstratos, e de alto nível, para implementações práticas e concretas. Floridi e Taddeo (2016) sugerem que esse tipo de ferramenta pode ser pensado como uma metodologia de diagnóstico, i.e., um modo para avaliar se um determinado modelo está alinhado com certos princípios éticos defendidos e definidos pelos desenvolvedores.

Iremos definir este tipo de ferramenta como "qualitativas", e nas próximas seções, apresentaremos as seguintes ferramentas de diagnóstico:

- *FAIR (Findability, Accessibility, Interoperability, and Reusability);*
- *Digital Catapult AI Ethics Framework;*
- *VCIO (Values, Criteria, Indicators, Observables).*

Construindo um Conjunto de Dados Justo (FAIR)

Quando os problemas de nosso modelo podem ser rastreados de volta ao conjunto de dados que utilizamos, uma possível solução é corrigir tal conjunto, de modo que sua distribuição de amostras, com respeito a atributos sensíveis, respeite alguma condição particular de justiça que desejamos implementar.

Por exemplo, FairFace[9] é um conjunto de dados de faces com uma distribuição balanceada entre gêneros, raças e idades, contendo 108,501 imagens. Kärkkäinen e Joo (2021) demonstraram que modelos treinados com FairFace são significativamente mais precisos do que outros modelos treinados com conjuntos como UTKFace, CelebA, LFWA+, mostrando performance consistente entre grupos (i.e., paridade preditiva entre raça, gênero e idade).

Para auxiliar desenvolvedores a identificar e escolher conjuntos de dados justos, podemos utilizar a metodologia proposta por Wilkinson et al. (2016): FAIR (Findability, Accessibility, Interoperability, and Reusability). A metodologia FAIR é uma metodologia para que desenvolvedores avaliem certas características do conjunto de dados que pretende utilizar, essas sendo: Capacidade de localização, Acessibilidade, Interoperabilidade e Reusabilidade.

Esses princípios servem para orientar desenvolvedores a averiguar três tipos de entidades: (1) dados (fontes digitais de informação); (2) metadados (informação sobre informação digital); e (3) infraestrutura (como os dados e metadados são estruturados e indexados).[10] Vejamos algumas das recomendações feitas pela metodologia FAIR.

---

[9] Disponível em: https://Github.com/joojs/fairface.

[10] Para mais informações sobre como implementar a metodologia FAIR, acesse https://www.go-fair.org/.

Capacidade de localização (Findability), i.e., dados e metadados devem ser de fácil acesso, tanto para humanos quanto computadores:

- Aos (meta)dados é atribuído um identificador global único e persistente (e.g., um repositório no GitHub, o "Orcid" do pesquisador responsável, o "Doi" de uma publicação que demonstra os resultados de aplicações do conjunto de dados);
- Os dados são descritos com ricos metadados (e.g., DICOM: *Digital Imaging and Communications in Medicine* é um protocolo para tratamento, armazenamento e transmissão de informação médicas em formato eletrônico, de modo a permitir que, por exemplo, informações de imagens médicas sejam acessíveis entre diferentes equipamentos de diagnóstico, geradores de imagens, computadores e hospitais);
- Os metadados incluem, de forma clara e explícita, o identificador dos dados que eles descrevem (e.g., a associação entre um arquivo de metadados e o conjunto de dados deve ser explicitamente mencionada nos metadados por um identificador globalmente único e persistente);
- Os (meta)dados são registrados ou indexados em um recurso pesquisável (e.g., o conjunto de dados pode ser encontrado por um motor de busca público, como, por exemplo, Google).

Acessibilidade (Accessibility), i.e., após que os (meta)dados sejam encontrados, os desenvolvedores devem saber quais os procedimentos para se ganhar acesso (e.g., autenticação e autorização) ao conjunto de dados:

- Os (meta)dados podem ser recuperados por seu identificador usando um protocolo de comunicação padronizado (e. g., o conjunto de dados pode ser acessado e baixado por um link http);

- O protocolo é aberto, livre, e universalmente implementável (e.g., o conjunto de dados é gratuito);
- O protocolo permite um procedimento de autenticação e autorização, quando necessário (e.g., conjuntos de dados devem ter suas condições de acesso explicitamente declaradas, como, por exemplo, autenticação por número de telefone);
- Metadados são acessíveis, mesmo quando os dados não estão mais disponíveis (e.g., se o conjunto de dados não pode mais ser acessado pelo seu identificador, metadados referentes a aquele conjunto irão explicitar que aquele conjunto de dados "expirou sua vida útil").

Interoperabilidade (Interoperability), i.e., o conjunto de dados deve estar em um formato que permita sua integração com diversas plataformas e aplicações:

- Os (meta)dados utilizam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento (e.g., o conjunto de dados está em JSON-LD);
- Os (meta)dados utilizam vocabulários que seguem os princípios FAIR;
- Os (meta)dados incluem referências qualificadas a outros (meta)dados (e.g., os metadados de um conjunto de dados podem referenciar outros conjuntos de dados similares).

Reusabilidade (Reusability), i.e., conjuntos de dados devem ser bem formatados, para que sua utilização possa ser replicada em diferentes situações:

- Os (meta)dados são ricamente descritos com uma pluralidade de atributos precisos e relevantes (e.g., além dos dados possuírem atributos claros e autoexplicativos, informações como "Para qual propósito os dados foram gerados/coletados?", possíveis vieses, se os dados são brutos ou processados, devem ser claramente explicitadas);

- Os (meta)dados são liberados com uma licença de uso de dados clara e acessível (e.g., o conjunto de dados é licenciado por uma licença MIT);
- (Meta)dados estão associados com a proveniência detalhada (e.g., os metadados contém uma página que descreve a história/origem do conjunto de dados);
- Os (meta)dados atendem às normas comunitárias relevantes para o domínio (e.g., conjuntos de dados devem ser formatados de modo padronizado, como, por exemplo, JSON-LD, de modo a permitir sua reusabilidade).

Caso você não utilize um conjunto de dados já pronto, será sua responsabilidade garantir que o conjunto de dados gerado para treinar o seu modelo siga estes critérios de boa conduta. Boa parte da engenharia de aprendizagem de máquina se resume a construir conjuntos de dados. Assim, durante esse processo, é importante tornar a descrição dos tipos de dados e atributos sendo coletados/utilizados a mais clara e detalhada possível. Segue abaixo algumas recomendações extras:

- Catalogue o número de amostras, para cada atributo sensível, que seu conjunto possui (e.g., quantas amostras são homens, quantas são mulheres, quantas amostras não têm o atributo de gênero declarado ou se declaram não-binários. Será que seu conjunto de dados permite que todo o espectro de gênero seja representado?);
- Descreva o domínio de origem dos seus dados (e.g., eles foram voluntariamente fornecidos? “Web crawling” para fins comerciais é permitido em seu país? Como a criação do seu conjunto de dados pode conflitar com o princípio de Privacidade?);
- Conheça intimamente o conjunto de dados, pois será da sua responsabilidade identificar possíveis vieses antes de sua fase de

implantação. Lembrando que nem todos os vieses são ruins, mas certos tipos de vieses podem sim gerar consequências indesejadas;

- Compartilhe seus achados. Se queremos desenvolver sistemas transparentes, projetos de código livre devem ser o padrão para a indústria da IA.

Outras ferramentas, como FAIR, podem ser encontradas na literatura. Gebru et al. (2018) fornecem uma ferramenta similar para avaliar conjuntos de dados utilizados para aprendizagem de máquina. De qualquer forma, e independente da ferramenta utilizada, é importante que conjuntos de dados sejam documentados e analisados de modo a evitar que consequências indesejadas venham a ocorrer após que tais modelos sejam implantados no mundo real.

Reportar os resultados de uma análise de ética e segurança é outro importante passo para desenvolvedores. Da mesma forma que medicamentos são vendidos com bulas contendo contra indicações, dosagens, efeitos colaterais, modelos de aprendizagem de máquina também devem ser apresentados de forma transparente.

Digital Catapult AI Ethics Framework

Desenvolvida pelo Comité de Ética da Digital Catapult,[11] o Digital Catapult AI Ethics Framework trata-se basicamente de uma metodologia em formato de entrevistas/questionários. Os questionários contemplam sete conceitos, onde cada conceito é explorado por questões específicas. Questões essas que visam explorar como uma organização está implementando preocupações éticas no desenvolvimento de seu produto.

A ideia por trás do Digital Catapult AI Ethics Framework, é que ao passar por este questionário, possíveis falhas de segurança ou certos tipos de má conduta serão melhor explicitados, e desenvolvedores tomaram ciência de sua existência. Os conceitos trabalhados por essa metodologia, como alguns exemplos de suas perguntas, são:[12]

Benefícios claros: os benefícios oferecidos por um produto devem ser claros e transparentes. Ao mesmo tempo, os benefícios devem superar os potenciais riscos associados ao produto desenvolvido.

- Quais são as metas, propósitos e aplicações pretendidas pelo produto desenvolvido?
- Quem ou o que pode se beneficiar do produto? Considere todos os grupos potenciais de beneficiários, sejam usuários individuais, grupos, ou a sociedade e o meio ambiente como um todo.

[11] O comitê de Ética da Digital Catapult busca traduzir teoria em Ética da IA para a prática. O comitê é presidido por Luciano Floridi, Professor de Filosofia e Ética da Informação da Universidade de Oxford, e diretor do Laboratório de Ética Digital da Universidade de Oxford. https://migarage.digicatapult.org.uk/ethics/ethics-committee/.

[12] Tal ferramenta, em sua versão completa, pode ser acessada através do seguinte endereço: https://migarage.digicatapult.org.uk/wp-content/uploads/2021/07/DC_AI_Ethics_Framework-2021.pdf.

- Tais benefícios podem mudar com o tempo?

Conhecer e gerenciar os riscos: os possíveis riscos associados com o uso indevido ou pretendido do produto, devem ser, dentro do possível, conhecidos pelos desenvolvedores.

- Foram considerados os riscos de outros usos previsíveis do produto, incluindo o uso indevido acidental ou malicioso do mesmo?
- Como podem ser comunicados os riscos potenciais ou riscos percebidos para os usuários, partes potencialmente afetadas, compradores ou comissionados?
- Todos os grupos potenciais em risco, sejam usuários individuais, grupos ou a sociedade e o meio ambiente como um todo, foram considerados?

Utilize os dados de forma responsável: o cumprimento da legislação vigente (e.g., LGPD - Lei nº 13.709/2018)[13], assim como outras ferramentas que auxiliem a garantir que dados sejam coletados e utilizados de forma ética (e.g., FAIR), são um ponto de partida básico para qualquer avaliação ética.

- Como os dados foram obtidos e como foi obtido o consentimento? Os dados são atuais?
- Os potenciais enviesamentos contidos nos dados foram examinados, bem compreendidos e documentados? Existe um plano para mitigar os mesmos?
- As pessoas podem se retirar do conjunto de dados? Tais pessoas também podem se remover de qualquer modelo resultante?

Ser digno de confiança: o ônus da prova de que seu produto é confiável e competente deve ser devidamente sustentado e provado pelos desenvolvedores. Este ônus também deve ser entregue em um formato

[13] Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm#art65.

que seja interpretável, de modo que usuários não sejam enganados ou confundidos.

- Dentro da empresa, há processos e ferramentas suficientes para garantir transparência, responsabilização, confiabilidade e adequação do produto desenvolvido?
- A natureza do produto é comunicada de forma que os usuários pretendidos, terceiros e o público em geral, possam acessar e entender?
- Quem é responsável se as coisas derem errado? Essas são as pessoas certas? Tais pessoas estão equipadas com as habilidades e conhecimentos necessários para assumir tal responsabilidade?

Diversidade, igualdade e inclusão: em um mundo plural e diverso, organizações que valorizam princípios como diversidade, igualdade e inclusão, devem ser o modelo a ser seguido. Assim, organizações devem estar aptas a prospectar as possíveis consequências da implantação de um produto para com uma vasta gama de grupos, com a intenção de que seu produto seja capaz de amenizar e mitigar as desigualdades e injustiças que se encontram estruturadas em nossa fábrica política, cultural e social.

- Existem processos para estabelecer se o produto pode ter um impacto negativo sobre os direitos e liberdades de indivíduos ou grupos?
- A organização tem uma política de diversidade e abrangência em relação ao recrutamento e retenção de pessoal?
- Onde a ética se encaixa nas práticas de contratação da empresa? Por exemplo, questões éticas são levantadas em entrevistas?

Comunicação transparente: a comunicação entre desenvolvedores (e a organização em geral) e usuários, partes potencialmente afetadas, investidores e comissionados, deve ser transparente, clara e inteligível.

Ao mesmo tempo, as vias de comunicação entre esses grupos devem permitir que preocupações e reclamações sejam tratadas de forma eficiente.

- A organização se comunica de forma direta, clara e honesta sobre quaisquer riscos potenciais do produto que está sendo fornecido?
- A empresa tem um sistema claro e fácil de usar para que as preocupações de terceiros/usuários, ou partes interessadas, sejam apontadas e tratadas?
- Existe alguma estratégia ou processo de comunicação se algo der errado (e.g., solicitação de devolução, "recall")?

Modelo de negócios: o conceito de "negociação justa" deve ser parte integrante da cultura organizacional de uma empresa, de modo que a maximização cega de capital não seja o único "guia normativo" que oriente e impulsione tal organização. Em outras palavras, organizações éticas também devem ser impulsionadas pela maximização do Bem Social.

- O impacto ambiental é considerado ao escolher fornecedores? Foram consideradas opções com fontes de energia limpa?
- Preços diferenciais foram considerados? Há alguma consideração ética em relação à estratégia de preços?
- Há algum grupo vulnerável que possa receber preços mais baixos?
- Existem dados que terceiros (e.g., instituições de caridade, pesquisadores) poderiam usar para benefício público?

A ideia por trás de uma entrevista realizada através do Digital Catapult AI Ethics Framework, é que problemas e fatos negligenciados sejam trazidos à luz do debate, e assim, medidas de mitigação possam começar a ser planejadas e elaboradas.

The AI Robotics Ethics Society®

VCIO (Valores, Critérios, Indicadores e Observáveis)

Krafft et al. (2020), através do AI Ethics Impact Group (liderado pela VDE Association for Electrical, Electronic & Information Technologies e a Bertelsmann Stiftung), propõe outro tipo de ferramenta translacional. Os autores apresentam o modelo VCIO (Values, Criteria, Indicators, Observables), algo que, de acordo com os autores, trata-se de uma abordagem única no campo da Ética da IA (Krafft et al., 2020, p. 6).

Assim como o Digital Catapult AI Ethics Framework, o modelo VCIO é uma forma de contextualizar preocupações éticas dentro do escopo de aplicação de um determinado modelo. O modelo VCIO é uma estrutura multimetodológica, onde sistemas de IA são: (1) avaliados em relação a uma série de princípios éticos pré-estabelecidos; (2) resultados são destilados em um selo de ética (AI Ethics Label); e por último (3), aplicações do modelo são classificadas através de uma matriz de risco.

VCIO é uma abordagem que busca identificar indicadores observáveis que possam servir como critérios de decisão para determinar se um princípio ético está sendo preservado ou não. A abordagem também busca esclarecer quando existem conflitos entre diferentes valores. Por exemplo, em certas aplicações (e.g., pesquisa médica), existe uma troca entre transparência e privacidade, onde é quase que impossível satisfazer ambos os lados (i.e., total transparência pode vir a significar pouca privacidade, e vice-versa).

Assim, a abordagem da VCIO opera em quatro níveis:

- *Valores:* aquilo que deve guiar as nossas ações;
- *Critérios:* aquilo que define se um valor (e.g., Justiça) foi violado ou não;
- *Indicadores:* como critérios (assim como valores) não podem ser diretamente observados, precisamos de indicadores que possam sinalizar se critérios estão sendo cumpridos;

- *Observáveis:* aspectos que podem ser observados e monitorados pelos indicadores.

De acordo com Krafft et al. (2020, p. 16):

> Contudo, não é possível derivar os níveis mais baixos [Indicadores e Observáveis] dos mais altos [Valores e Critérios] de uma forma direta, ou seja, dedutiva. Em vez disso, a carga normativa percorre os quatro níveis e requer novas deliberações em todos os níveis, no decorrer das quais as instâncias particulares devem ser negociadas em detalhes. [...] Como não há relações dedutivas entre valores, critérios, indicadores e observáveis [...] decisões normativas devem ser tomadas em um contexto científico e tecnicamente informado.

Como um exemplo, se determinarmos "Paridade Preditiva" como um critério para "Justiça", podemos utilizar a precisão de um modelo como indicador, e monitorar esta métrica de performance com relação a diferentes grupos (quantidade observável). Se determinarmos "Sustentabilidade" como um valor, podemos utilizar "pegada de carbono" como um critério, monitorando, por exemplo, a emissão de carbono gerada para se treinar um modelo específico. Ou podemos monitorar se uma organização em particular opta por fontes de energia limpa para treinar seus modelos e rodar seus servidores.

Já que não existe uma forma clara e objetiva de se determinar critérios, indicadores e observáveis dos valores escolhidos (podemos até mesmo afirmar que a escolha de todos estes será uma escolha normativa por natureza), o ônus de se provar que existe uma correlação entre aquilo que se defende e aquilo que é monitorado cai sobre os desenvolvedores.

Caso existam (e geralmente existem) conflitos entre valores, desenvolvedores podem hierarquizá-los de forma a priorizar, dependendo do contexto em que um modelo será aplicado, diferentes valores. Por exemplo, desenvolvedores podem optar por priorizar valores cujos

critérios, indicadores e observáveis sejam mais claros de serem monitorados e quantificados. Se um valor não possui nenhuma forma clara de ser monitorado (e.g., Responsabilização), esse pode ser utilizado como critério de desempate entre dois valores conflitantes (e.g., seja por quebra de privacidade ou por falta de transparência, por qual dessas violações será mais fácil atribuir responsabilização aos devidos responsáveis? Qual violação pode gerar maiores danos aos envolvidos?).

Na tabela abaixo, vemos a aplicação do modelo VCIO na análise do princípio de Justiça.

Valores	Justiça		
Critérios	Avaliação de diferentes fontes de possíveis vieses para garantir Equidade/Justiça.		
Indicadores	Os dados de treinamento foram analisados para se identificar possíveis vieses?	Os procedimentos de rotulagem de dados foram avaliados?	O modelo possui paridade preditiva entre diferentes grupos demográficos?
Observáveis	Sim, todos os potenciais vieses do modelo foram reportados.	Sim, a rotulagem dos dados foi vistoriada por avaliadores externos.	Sim, paridade preditiva é garantida.
	Apenas alguns vieses são de conhecimento dos desenvolvedores.	Sim, a rotulagem dos dados foi vistoriada por avaliadores internos.	Paridade preditiva é garantida apenas dentro de um percentual de erro pré-determinado.
	Não.	Não.	Não.

Esta tabela foi adaptada e modificada de um dos exemplos fornecidos por Krafft et al. (2020, p. 22).

Tabelas como essa podem ser aplicadas para contemplar uma série de valores diferentes, onde para cada valor podemos atribuir mais de um critério, cada um com seus respectivos indicadores e observáveis (a tabela acima é apenas um exemplo simplificado). Assim, a ideia principal do modelo VCIO é hierarquizar valores, critérios, indicadores e observáveis de modo que conceitos abstratos (e.g., Justiça) possam ser ancorados em variáveis observáveis (e.g., as taxas de precisão são equivalentes para todos os grupos considerados pelo modelo dentro de um limite de erro pré-estabelecido como aceitável).

Para facilitar a interpretabilidade da análise proposta pelo modelo VCIO, os resultados são então condensados em um "Selo de Ética", i.e., um indicador que seja de fácil entendimento para cidadãos, usuários, consumidores, legisladores ou órgãos de regulamentação.

O selo proposto por Krafft et al. (2020) inclui uma classificação para cada um dos valores contemplados por uma análise ética. No exemplo abaixo, os valores utilizados são *Transparência*, *Responsabilização*, *Privacidade*,

Justiça, *Confiabilidade* e *Sustentabilidade*. Contudo, o modelo certamente pode ser estendido para contemplar outros valores.

A classificação sugerida é feita por letras, de A à G (sete níveis), onde "A" é a classificação mais alta (e.g., um modelo com pontuação "A" em Justiça é um modelo onde todos, ou a maioria, dos critérios são atendidos pelas medidas observáveis mais rigorosas). Krafft et al. (2020) sugerem sete níveis para que a granularidade dos observáveis seja melhor expressada (na tabela acima, utilizamos apenas três).

Selo de Ética em IA (Krafft et al., 2020, p. 13)

Se escolhêssemos utilizar a tabela exemplo deste Guia, poderíamos escolher criar uma classificação com apenas três níveis ("A", "B" e "C", ou "Verde", "Amarelo", "Vermelho"). Cada observável corresponderia a uma classificação (e.g., "Verde = A"), e a nota final atribuída a um sistema de IA seria feita pela agregação das diferentes classificações observáveis. Por exemplo, um modelo pode receber uma classificação "B" em Justiça, caso ele contemple dois indicadores com observáveis "A" e um indicador com um observável "C".[14]

Após que os níveis de classificação sejam definidos, assim como a granularidade dos observáveis, ainda é preciso definir uma forma de agregar tais pontuações. Existem diversas formas de se fazer um

[14] Todos os vieses do modelo são conhecidos/explicitados (A); o modelo garante paridade preditiva (A); contudo, o procedimento de rotulagem dos dados utilizado para o treinamento do modelo não foi auditado, nem por avaliadores internos ou externos (C) (i.e., "A + A + C = B").

procedimento de agregação desse tipo, e Krafft et al. (2020) sugerem métodos como média aritmética, média harmónica, e até mesmo a definição de critérios mínimos para que certas classificações sejam alcançadas.[15]

Agora, um outro passo que precisamos dar em uma análise ética é avaliar o contexto, e os potenciais riscos associados, de uma aplicação. Como foi dito, existe pouca (ou nenhuma) análise ética necessária na implementação de um modelo de IA criado para monitoramento de processos industriais.[16] Contudo, existem contextos de aplicação onde sistemas de IA, de um ponto de vista ético, não deveriam jamais ser utilizados. Por exemplo, uma decisão de alto risco, como se devemos ou não desligar equipamentos de suporte de vida de um paciente com morte cerebral, não deveria (a princípio) ser tomada por um modelo de inferência estatística. Ou, uma pena capital (i.e., "pena de morte"), jamais deveria ser prescrita e sentenciada por um sistema de IA.

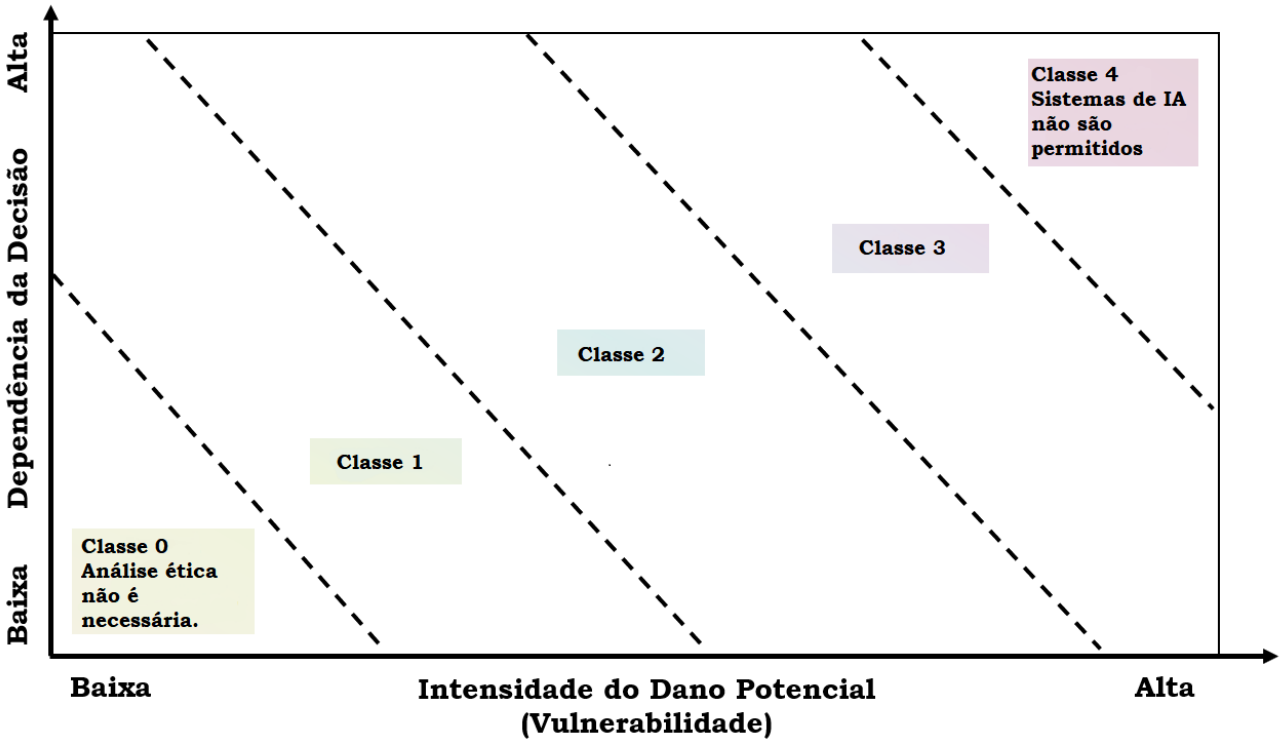
Dessa forma, ao analisarmos o contexto de uma aplicação, esse pode nos informar também o rigor de nossa avaliação (e.g., quais observáveis devem ser cumpridos para que um modelo receba uma classificação mínima), se a nossa avaliação é necessária (e.g., o modelo não necessita nenhuma análise ética), ou se o modelo não deve ser implantado de forma alguma (e.g., o resultado da análise ética recomenda a não permitir a implantação de certa aplicação).

Krafft e Zweig (2019) sugerem que os riscos de uma determinada aplicação sejam analisados em uma matriz de risco bidimensional. Matrizes de risco são comumente utilizadas para a avaliação do nível de

[15] Os detalhes e nuances do modelo VCIO podem ser todos encontrados na publicação original de Krafft et al. (2020).

[16] Talvez seja necessário, caso tal modelo venha a causar um certo deslocamento de mão de obra, i.e., pessoas perderão seus empregos para sistemas de IA.

risco de um sistema. A matriz proposta por Krafft e Zweig (2019) possui dois fatores: (1) a intensidade do dano potencial; e (2) a dependência da(s) pessoa(s) afetada(s) pelo respectivo modelo. Os autores dividem sua matriz de risco em cinco classes, de 0 ("nenhum risco") a 4 ("a implantação do modelo não deve ser permitida").[17]



Matriz de Risco (Krafft & Zweig, 2019, p. 32).

Nesta matriz, o risco é definido por dois eixos diferentes. O eixo vertical representa o quanto as decisões de um modelo (ADM – "algorithmic decision-making systems") poderiam vir a afetar pessoas. Abaixo temos algumas perguntas que podem nos guiar para a avaliação desta dimensão:

[17] De acordo com Krafft & Zweig (2019), o modelo de análise de risco não foi concebido de forma a classificar todos os possíveis contextos de risco exaustivamente. Certamente que podemos aumentar a granularidade do espectro de risco, contudo, a ideia principal por trás dessa ferramenta ainda é significativa, i.e., que o risco de uma determinada aplicação seja avaliada antes de sua implantação.

- *Qual é a função que o sistema está automatizando (e.g., discernindo gatos de cachorros, ou determinando o desligamento emergencial de uma usina nuclear)?*
- *Existe algum tipo de supervisão humana?*
- *Se o sistema apresentar alguma falha em seu funcionamento, como isso pode afetar as partes envolvidas?*
- *As decisões do sistema podem ser contestadas? De que forma?*

Enquanto que o eixo horizontal representa a intensidade do dano potencial causado pelas saídas do modelo:

- *O modelo do pode vir a impactar direitos humanos fundamentais?*
- *Qual o nível desse impacto (e.g., perda de um benefício, dano físico, perda de uma vida)?*
- *Os impactos serão voltados para pessoas físicas? Pessoas jurídicas? Indivíduos ou Organizações?*

Como exemplos de modelos de IA para cada classe, podemos citar:

- *Classe 0:* sistemas para automação de processos industriais, sistemas para automação de previsões meteorológicas, sistemas de recomendação de produtos;
- *Classe 1:* sistemas de recomendação para buscas personalizadas em motores de busca, sistemas de recomendação em redes sociais, sistemas de recomendação em plataformas de streaming;
- *Classe 2:* sistemas de recomendação personalizada para empregos, sistemas de recomendação personalizada para serviços, modelos de linguagem para conversação (i.e., chatbots);
- *Classe 3:* sistemas de recomendação para propagandas eleitorais, sistemas de visão computacional para aplicação da lei, sistemas de avaliação de reincidência criminal, sistemas de avaliação de escore de crédito, veículos autônomos;

- *Classe 4:* armas autônomas, juízes autônomos.

Outro exemplo de matriz de risco, é a matriz de risco da MIL-STD-882E (Military Standard 882, Department of Defense Standard Practice System Safety, EUA).[18] A matriz de risco MIL-STD-882E para avaliações qualitativas tem duas categorias de avaliação: *Severidade* e *Probabilidade*.

Matriz de Avaliação de Risco				
Probabilidade/ Severidade	**Catastrófico (1)**	**Crítico (2)**	**Marginal (3)**	**Negligível (4)**
Frequente (A)	**Alto**	**Alto**	**Sério**	**Médio**
Provável (B)	**Alto**	**Alto**	**Sério**	**Médio**
Ocasional (C)	**Alto**	**Sério**	**Médio**	**Baixo**
Remoto (D)	**Sério**	**Médio**	**Médio**	**Baixo**
Improvável (E)	**Médio**	**Médio**	**Médio**	**Baixo**
Eliminado (F)	**Eliminado**			

Severidade pode ser definida pelo seguinte conjunto de categorias:

- *Catastrófico:* risco de morte (e.g., armas autônomas atacando civis);
- *Crítico:* risco de lesões graves (e.g., acidentes de trânsito causados por veículos autônomos);
- *Marginais:* danos/lesões menores (e.g., classificações incorretas/prejudiciais geradas por um ADM);

[18] MIL-STD-882E, Department of Defense Standard Practice: System Safety (11 de maio de 2012). Disponível em: http://everyspec.com/MIL-STD/MIL-STD-0800-0899/MIL-STD-882E_41682/.

- *Negligenciável:* danos/lesões negligenciáveis (e.g., seu feed de vídeos não contém suas séries favoritas).

Enquanto isso, Probabilidade é a estimativa da frequência de um evento que pode vir a acontecer no futuro (algo que muitas vezes é difícil, ou impossível, de ser determinado com precisão):

- *Frequente:* evento que pode ocorrer com frequência (e.g., uma classificação errada a cada 10 amostras);
- *Provável:* ocorrerá várias vezes na vida do sistema (e.g., uma classificação errada a cada 100 amostras);
- *Ocasional:* eventos que podem vir a ocorrer em algum momento da vida do sistema (e.g., uma classificação errada a cada 1,000 amostras);
- *Remoto:* evento improvável que ocorra, mas ainda pode ocorrer (e.g., uma classificação errada a cada 10,000);
- *Improvável:* evento extremamente improvável de ocorrer (e.g., uma classificação errada a cada 100,000);
- *Impossível:* Igual a uma probabilidade de zero.

Certamente que os exemplos citados acima podem ser contestados. Dado a natureza ambígua e contexto dependente da Ética/Segurança quando aplicada a situações complexas do mundo real, argumentos podem ser feitos para se definir a qual classe uma aplicação "realmente" pertence, ou qual o "verdadeiro" nível de severidade/probabilidade[19] de um sistema de IA falhando em agir de forma segura. Contudo, acredito que o importante não seja exatamente o resultado final (i.e., a classificação

[19] Uma classificação errada para cada 1,000 amostras pode parecer pouco, mas caso a aplicação sendo avaliada realize uma chamada por segundo ao modelo, e o modelo opere durante 4 horas/dia, isso são 14400 chamadas ao modelo por dia (~15 erros por dia). Dependendo da aplicação, isso pode vir a ser considerado como um risco alto.

“exata” de uma aplicação) mas sim o processo de deliberação que levará a esse resultado (i.e., a análise Ética em si).

Ao mesmo tempo, de modo a otimizar os processos de regulamentação de sistemas de IA, uma análise de risco (e.g., por uma divisão de classes de risco) pode auxiliar a definir qual o rigor que devemos ter em nossa avaliação. Dessa forma, aplicações que se enquadrem em classes de risco diferentes devem ser abordadas de forma diferente.

Por exemplo, podemos vir (como sociedade) a definir que enquanto aplicações que envolvam baixo risco (e.g., classes 0, 1 e 2, ou níveis de risco Eliminado, Baixo e Médio) podem ser auditadas internamente (i.e., pela própria organização), enquanto que aplicações de alto risco (e.g., classes 3 e 4, ou Sério e Alto) devem ser auditadas também por órgãos reguladores externos (e.g., o governo, a ACM, a IEEE). Também podemos definir que certas classes de aplicação, para que sistemas de IA possam ser seguramente implantados, obtenham valores mínimos em sua avaliação (e.g., todas as aplicações que se enquadrem na Classe 2 devem obter uma avaliação “B” em todos os valores avaliados).

Morley et al (2021, p. 250) resumem o conceito de “Ética como um Serviço” em dois tipos de “esferas de responsabilização”, que sintetizam as preocupações levantadas pelas ferramentas translacionais apresentadas nesta seção:

- *Responsabilidade Interna:* Definir contextualmente o significado de cada princípio ético explicitado por um Código de Ética criado por órgãos regulamentadores (i.e., responsabilidade externa). Selecionar a utilização de ferramentas/métodos a partir de uma lista pré-aprovada de ferramentas/métodos disponíveis. Conduzir a revisão ética do próprio produto em todas as etapas de desenvolvimento e implementação, incluindo uma análise ética prospectiva para o futuro.
- *Responsabilidade Externa:* Desenvolver um Código de Ética, revisá-lo regularmente, e desenvolver um processo que desenvolvedores

de IA devam seguir para aplicar contextualmente os princípios éticos definidos por tal código. Avaliar as ferramentas/métodos disponíveis, e compilar uma lista pré-aprovada de tais ferramentas para que desenvolvedores possam utilizá-las no desenvolvimento de seus produtos. Auditar os sistemas de IA desenvolvidos para averiguar sua conformidade para com o Código de Ética vigente (e.g., o Ethically Aligned Design da IEEE; o Projeto de Lei 21/2020;[20] O Código de Ética da ACM[21]).

Distribuir a responsabilização da governança da IA (e sua operacionalização Ética) dessa forma, garante uma forma relativamente clara de qual o papel dos diferentes atores dessa hierarquia de serviços. Seja um membro do comitê de ética da IEEE, trabalhando para a atualização do atual Código de Ética vigente, ou um engenheiro de segurança de uma empresa realizando o diagnóstico/avaliação de um modelo, cada ator tem seu papel a cumprir.

No fim, as ferramentas translacionais apresentadas podem ser utilizadas de forma individual ou em conjunto. Tais ferramentas permitem uma abordagem geral para se implementar a ética no desenvolvimento de sistemas inteligentes:

- Uma organização que planeja desenvolver um sistema de IA para uma aplicação específica, pode usar questionário/lista de verificação inicial (e.g., Digital Catapult AI Ethics Framework) para determinar o risco ético de uma aplicação. Dependendo do risco envolvido (e.g., matriz de risco do modelo VCIO classifica a

[20] Projeto de lei que estabelece os fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil. Disponível em: https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=1853928.

[21] Código de Ética e Conduta Profissional da ACM (Association for Computing Machinery). Disponível em: https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-and-professional-conduct.pdf.

aplicação como “Classe 0”), o processo termina nesta etapa. Caso houver questões éticas para se considerar, então a organização realiza uma avaliação completa de sua aplicação;

- Uma avaliação completa deve contemplar todas as etapas de desenvolvimento de um sistema de IA (i.e., Coleta de dados, Desenvolvimento do Modelo, Avaliação do Modelo, Pós-processamento, Implantação, Monitoramento). Cada etapa pode ser comprometida por diferentes fontes de problemas (e.g., os dados coletados são enviesados por vieses históricos). Ferramentas como FAIR, a Digital Catapult AI Ethics Framework, o Modelo VCIO, entre outras, podem auxiliar desenvolvedores a tornar tais problemas mais evidentes;
- Cada contexto de aplicação possui suas próprias especificidades. Certos valores e princípios éticos podem não fazer sentido para uma determinada aplicação. Cabe aos desenvolvedores (e órgãos de regulamentação) acessar quais princípios éticos devem ser priorizados em um determinado contexto de aplicação. Princípios éticos devem ser aterrados a quantidades observáveis e verificáveis, de forma que uma avaliação ética possa se basear em certos critérios objetivos de avaliação (e.g., o Modelo VCIO realiza isso com a metodologia “Valores, Critérios, Indicadores e Observáveis”);
- Dependendo do contexto, certos valores encontraram-se em oposição (e.g., Transparência e Privacidade), enquanto que outros valores poderão apenas ser aproximados dentro de um contexto de aplicabilidade (e.g., Justiça algorítmica “total” sofre de um teorema de impossibilidade). Cabe aos desenvolvedores explicitar, de forma transparente, os comprometimentos e compromissos realizados;
- Este processo de avaliação, dependendo do risco envolvido pela aplicação (e.g., matriz de risco do modelo VCIO, matriz de risco da MIL-STD-882 E) pode ser diretamente auditado pela organização, ou necessitar de uma avaliação externa, realizada pelos órgãos regulamentadores responsáveis;

- Após o final de uma avaliação, os resultados devem ser apresentados de forma clara e transparente para todas as partes envolvidas na utilização do sistema desenvolvido (e.g., Selo de Ética do Modelo VCIO).

Contudo, devemos lembrar que certos problemas podem apenas surgir após a fase de implantação de um modelo. Assim, não podemos reduzir a análise ética e a engenharia de segurança a apenas *"listas de verificação a serem preenchidas".*

Seriam ferramentas translacionais o bastante?

Por mais que ferramentas translacionais auxiliem a aproximar a teoria ética à prática do desenvolvimento de sistemas inteligentes, é importante que tenhamos consciência de que tais estratégias, sozinhas, não garantem que um determinado modelo/produto não irá gerar consequências indesejadas.

Os atores responsáveis por administrar uma avaliação ética, com suas próprias noções éticas particulares, podem nem sempre estarem alinhados com o "Bem Social" (Green, 2019, Krishnan, 2019). Assim, é necessário que exista um esforço para se alinhar tais visões. Ou seja, é necessário que os desenvolvedores tenham um entendimento do que significa o "Bem Social". Por isso que comissões de avaliação ética devem sempre ser formadas por um grupo interdisciplinar com membros de diversas áreas do conhecimento (e.g., engenheiros, cientistas da computação, filósofos, sociólogos, advogados, etc.).

Uma crítica levantada contra ferramentas translacionais, é que tais métodos são "extra-empíricos" (Fazelpour & Lipton, 2020). Ou seja, enquanto que tais ferramentas procuram uma base empírica e objetiva para testar e avaliar noções de ética no desenvolvimento de sistemas inteligentes, essas próprias ferramentas não são "em si" sujeitas a uma avaliação "empírica e objetiva". Algo que, como Morley et al. (2021) apontam, torna tais metodologias sujeitas a manipulação daqueles que as aplicam.

Uma avaliação ética não pode ser reduzida a apenas um teste "único" ou um inventário a ser preenchido. O papel do engenheiro de segurança em ética da IA é um processo constante, pois modelos devem ser monitorados constantemente. Sem uma manutenção constante desses modelos, ferramentas translacionais não garantem que um sistema de IA será benéfico ou seguro. Imagine uma empresa de elevadores onde não se realiza uma rotina periódica de avaliação e inspeção de seus produtos, e

apenas vende-se elevadores com um selo dizendo “100% seguro”. Você compraria (ou usaria) um dos elevadores desta empresa?

Na verdade, você nem poderia (legalmente) comprar um produto desses, pois na maioria dos países, empresas que não implementam um “Programa de Manutenção Preventiva” para com esse tipo de tecnologia nem mesmo podem prestar serviços de forma legal. [22]

Da mesma forma que esse tipo de implementação já é um procedimento de “praxe” para tecnologias como elevadores, o mesmo deve se tornar rotineiro para com a manutenção de sistemas de IA. Sistemas inteligentes não podem ser produzidos, implementados, e depois abandonados por seus desenvolvedores. E é isto que se espera de uma organização que realmente busque desenvolver inteligência artificial ética e segura.

Para que isso seja alcançado, a Ética não pode ser totalmente reduzida a procedimentos de diagnóstico e avaliação, mas sim tratada como um *serviço preventivo* que deve ser regularmente empregado.

A partir da próxima seção, veremos como questões de segurança vêm sendo abordadas pela literatura e setor privado, e como podemos incrementar as metodologias qualitativas até agora apresentadas com ferramentas mais quantitativas.

[22] Portaria SIT Nº 224 DE 06/05/2011. Disponível em: https://www.legisweb.com.br/legislacao/?id=232119.

Segurança da IA

Segurança da IA (AI Safety) é, em si, sua própria área de pesquisa, com suas próprias preocupações. Essa área surgiu da necessidade de se desenvolver métodos para lidar com sistemas opacos, complexos, frágeis quando operando fora de sua distribuição, não moduláveis e dificilmente interpretáveis. E tais sistemas necessitam de sua própria forma de tratamento especial:

> Assim como, historicamente, as metodologias de segurança desenvolvidas para hardware eletromecânico não generalizaram bem para as novas questões levantadas pelo software, devemos esperar que as metodologias de segurança de software não irão generalizar bem para as novas complexidades e perigos da Aprendizagem de Máquina (Hendrycks et al., 2021a, p. 2).

Jurić et al. (2020), em sua revisão bibliométrica da literatura em segurança da IA, sugerem que os principais tópicos sendo trabalhados na área são:

- *Interpretabilidade:* Como interpretar a tomada de decisão de algoritmos opacos, como redes neurais profundas (Guidotti et al., 2018)? Ao mesmo tempo, como interpretar os resultados de nossas próprias ferramentas de interpretabilidade?
- *Corrigibilidade:* Como tornar agentes potencialmente falhos, mesmo que agentes racionais (maximizadores de utilidade esperada) possuem um forte incentivo instrumental para preservar seus objetivos terminais, corrigíveis (Soares et al., 2015)?
- *Robustez a Ataques Adversariais:* Redes neurais são altamente suscetíveis a ataques adversariais, i.e., ataques especialmente desenhados para enganá-las (Yuan et al., 2019). Como podemos proteger nossos sistemas contra esta forma de ataque?

- *Exploração Segura e Mudança de Domínio:* Geralmente, o domínio de treinamento não é uma representação perfeita do domínio real onde o agente irá operar. Como podemos garantir o comportamento "seguro" de nossos modelos quando operando em domínios muito diferentes daqueles que foram vistos em seu treinamento (Amodei et al., 2016)?
- *Aprendizagem de Valores e Especificação de Objetivos:* Conforme buscamos integrar sistemas de IA em ambientes cada vez mais complexos, as tarefas que esperamos que tais sistemas resolvam também se tornam mais complexas. Especificar uma função objetiva a ser otimizada por um sistema de IA de forma "clara" (i.e., sem erros de especificação) não é uma tarefa simples, já que valores e preferências humanas podem ser extremamente difíceis de especificar (Soares, 2016).

Já Hendrycks et al. (2021a) apresentam os seguintes problemas técnicos que encontramos em aprendizagem de máquina. Problemas esses que tendem a se tornar gradualmente mais proeminentes conforme modelos são implementados em aplicações cada vez mais complexas e de alto risco:

- *Robustez:* A criação de modelos que sejam resistentes a ataques adversariais e situações incomuns (i.e., situações fora de sua distribuição de treinamento). Atualmente, modelos treinados por aprendizagem de máquina ainda são frágeis e rígidos, não operando bem ambientes dinâmicos e mutáveis. Em um mundo repleto de eventos raros acontecendo a todo momento, tais modelos devem ser extremamente robustos;
- *Monitoramento:* A detecção de utilização maliciosa, mal funcionamento ou funcionalidades não pretendidas. Da mesma forma que usinas nucleares são monitoradas por HROs (high-

reliability organizations), sistemas de aprendizagem de máquina futuros podem vir a ser monitorados da mesma forma (e.g., sistemas inteligentes de gestão de tráfego controlando cidades povoadas por automóveis autônomos). Dessa forma, se torna necessário desenvolver metodologias para auxiliar o monitoramento e supervisão desse tipo de sistemas;

- *Alinhamento:* Criação de modelos que otimizam, de forma robusta, objetivos difíceis de serem especificados (e.g., valores humanos). Sistemas de IA muitas vezes apresentam um certo nível de agência (e.g., possuem e otimizam objetivos). Algo que difere tais sistemas de outras formas de tecnologia. Idealmente, gostaríamos de criar agentes que "prefiram" bons estados-de-mundo. Entretanto, *o que define um "bom estado-de-mundo"?* Proxies de objetivos podem ser: (1) difíceis de se especificar; (2) difíceis de se otimizar; (3) frágeis; e (4) estimular comportamentos indesejados (e.g., hackeamento de recompensa);
- *Segurança Externa:* Modelos podem estar integrados em ambientes inseguros, como software malfeito e organizações mal estruturadas. Dada a fragilidade que modelos treinados por aprendizagem de máquina apresentam, é importante tornar seus ambientes de implantação seguros, seja desenvolvendo softwares resilientes a ciberataques, ou criando políticas de governança que visem tornar a implantação de tais modelos segura.

É importante ressaltar que todas as vias de pesquisa citadas, com suas problemáticas particulares, permanecem sendo problemas em aberto em segurança da IA (e do próprio campo da Aprendizagem de Máquina).[23]

Como qualquer campo de pesquisa emergente, as preocupações e contribuições provenientes da área de Segurança da IA ainda não penetraram o "mainstream" da indústria e academia. Por exemplo, se formos analisar os principais projetos de pesquisa e desenvolvimento

[23] Para os interessados, Critch & Krueger (2020) apresentam uma extensa análise, com diversas sugestões e vias de pesquisa, do campo de segurança da IA.

(P&D) de IA avançada (i.e., projetos que buscam avançar o estado-da-arte do campo), vemos que apenas uma pequena minoria realiza algum tipo de pesquisa voltada para a área de segurança.

Em 2017, Baum (2017) identificou 45 projetos de P&D com os objetivos de desenvolver IA avançada. Dos 45 projetos revisados, apenas 13 possuíam envolvimento ativo com a área de segurança, enquanto que a grande maioria não especificava nenhum tipo de pesquisa voltada para a área de segurança da IA. Fitzgerald et al. (2020) atualizaram os achados de Baum (2017), aumentando a contagem de projetos para 72 projetos de P&D em 2020, ativos, focados em desenvolver IA avançada. Dos 72 projetos listados, apenas 18 possuem engajamento ativo com a área de segurança da IA.

Produzimos uma tabela/resumo dos achados de Fitzgerald et al. (2020), "2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy", um trabalho comissionado pelo Global Catastrophic Risk Institute. Nesta tabela se encontram: (1) o nome do Projeto (com link para sua página); (2) o país/líder que o sedia; (3) a instituição (e tipo de instituição) responsável pelo projeto; (4) se tal projeto possui vínculos com o setor Militar; (5) se o projeto é de Código aberto; (6) o tamanho do projeto; e por fim (7) o engajamento com a área de Segurança da IA de cada projeto. A tabela pode ser encontrada no link citado no rodapé.[24]

Por esses resultados, podemos ver que, como fora dito no início desta seção, a segurança da IA ainda é "algo novo a ser integrado". Contudo, temos bons exemplos de organizações que investem e se preocupam com

[24] Vigilância de Segurança da IA: P&D para Inteligência Artificial Avançada (2020). Disponível em: https://www.airespucrs.org/post/vigil%C3%A2ncia-de-seguran%C3%A7a-da-ia-p-d-para-intelig%C3%AAncia-artificial-avan%C3%A7ada-2020.

o desenvolvimento ético e seguro de suas aplicações. Tomemos como exemplo duas das maiores organizações envolvidas no desenvolvimento de IA: DeepMind[25] e OpenAI.[26]

DeepMind é um projeto da Google sediado em Londres (RU) é liderada por Demis Hassabis e Shane Legg. De seus laboratórios, além de alguns dos mais proficientes e gerais modelos de IA já produzidos (Mnhi et al., 2013; Silver et al., 2016; Badia et al., 2020), muitos estudos relacionados à Segurança da IA tem sido produzidos e publicados (Leike et al., 2017; Everitt et al., 2019; Mikulik et al., 2020; Kenton et al., 2021). DeepMind também colabora com OpenAI em projetos voltados à segurança da IA.

Já a OpenAI, uma organização sem fins lucrativos de pesquisa em IA, também é responsável por empurrar o estado-da-arte em diversas áreas do campo (Brown et al., 2020; Chen et al., 2021), publicando a maior parte de seus achados de forma aberta (open-source), e promovendo abertamente sua missão de "*tentar construir diretamente uma IAG segura e benéfica*".[27]

Tomemos como exemplo dois dos modelos mais recentes lançados pela OpenAI: GPT-3 e Codex.[28]

GPT-3 (Generative Pre-Train Transformer 3), um Transformer com 175 bilhões de parâmetros, é um modelo de aprendizagem de máquina treinado de forma não-supervisionada (Self-Supervised) capaz de gerar amostras de textos como poemas, artigos, notícias, além de resolver diversos problemas ligados à área de NLP, sem precisar de nenhum tipo de pós-processamento ou afinação. Contudo, qual o tipo de

[25] https://deepmind.com/.

[26] https://openai.com/.

[27] https://openai.com/about/.

[28] Esses modelos ainda não foram (por motivos de segurança) lançados abertamente ao público. Contudo, as publicações de Brow et al. (2020) e Chen et al. (2021) descrevem o processo de treinamento de tais modelos. Os modelos também podem ser acessados via API pela plataforma OpenAI beta, disponível em: https://beta.openai.com/.

comportamento indesejado que podemos esperar de um modelo com este interagindo com o mundo real?

Em sua publicação, Brown et al. (2020) realizam uma extensa análise de segurança do modelo desenvolvido. Nela, os autores relatam potenciais aplicações maliciosas (e.g., desinformação, spam, crimes cibernéticos), problemas relacionados com a equidade, preconceito e representatividade (e.g., gênero, raça, religião) e até mesmo o consumo de energia relacionado a utilização do modelo (i.e., Sustentabilidade).

Já Codex trata-se de um modelo capaz de transcopilar comandos dados em linguagem natural para código (e.g., Python). Codex foi treinado a partir de modelos de linguagem GPT afinados com conjuntos de dados públicos de código aberto (GitHub). Vejamos um exemplo gerado pelo API da OpenAI beta:

Comando:

"""

1. Create a function that counts prime numbers between 0 and 100.
2. Plot the number of prime numbers between 0 and 100 as a step function.
3. Label the plot “Number of Primes between 1 and 100”.
4. Use sympy, numpy, and matplot as modules.

"""

Output:

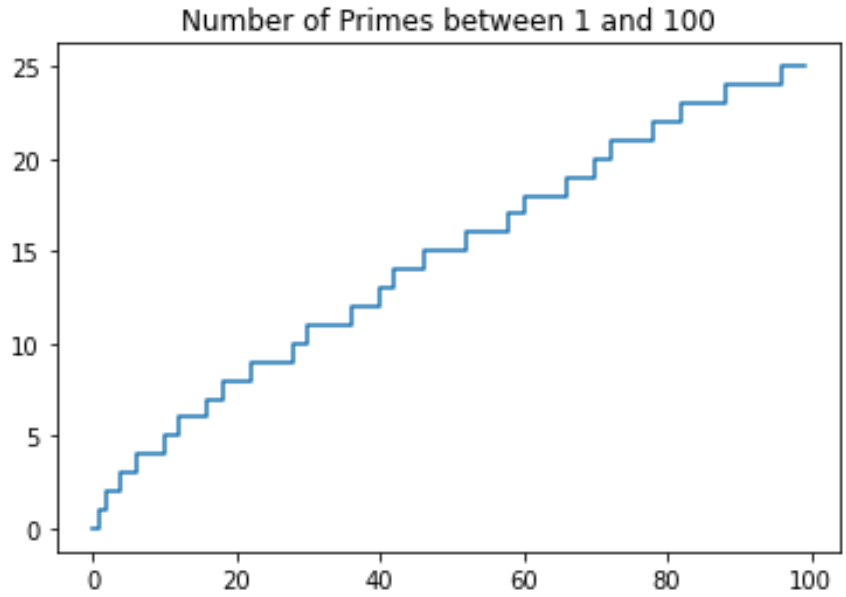
```
import sympy
import numpy as np
import matplotlib.pyplot as plt

def prime_count(n):
    count = 0
    for i in range(2, n+1):
        if sympy.isprime(i):
            count += 1
    return count
```

```
x = np.arange(0, 100, 1)
y = np.array([prime_count(i) for i in x])

plt.step(x, y)
plt.title("Number of Primes between 1 and 100")
plt.show()
```

Você pode testar esse resultado você mesmo. Apenas copie e cole o código acima em um Google Colab notebook[29] e pressione "play". O programa produzirá o que foi pedido (e acerta o número, i.e., 25):



Novamente, podemos ver na publicação de Chen et al. (2021) uma extensa análise das implicações e possíveis impactos deste tipo de tecnologia, algo que promete tornar a capacidade de escrever e gerar código algo extremamente acessível para todos aqueles que saibam ler e escrever comandos em língua inglesa (ou sejam letrados e tenham acesso a um tradutor).

Em sua publicação, Chen et al. (2021) mencionam fatores de risco como:

- O Codex nem sempre produz códigos que estão alinhados com a intenção do programador. Aqui, definimos desalinhamento (ou

[29] https://colab.research.google.com/.

"falha de alinhamento") como "*quando o sistema é designado para realizar alguma tarefa X, e o sistema é capaz de realizar X, mas "opta" por não o fazer*" (Chen et al., 2021, p. 11). Isto contrasta com a situação em que um sistema não faz X porque não tem a capacidade de fazer X, i.e., o sistema é apenas incompetente;

- O Codex pode sugerir soluções que superficialmente pareçam corretas, mas que na verdade não realizam a tarefa pretendida pelo usuário (i.e., confiança excessiva);
- Como no caso de outros modelos de linguagem, Codex pode ser solicitado de forma a produzir código/comentários que contêm conteúdo racista e denegridor;
- Os autores avaliaram os impactos econômicos no mercado de trabalho que modelos de geração automática de código podem causar (e.g., reduzindo o valor do trabalho de engenheiros e desenvolvedores de software);
- Os autores avaliaram a probabilidade de modelos de geração automática de código auxiliarem na criação de malware (assistindo na realização de cibercrimes);
- Os autores avaliaram qual o impacto ambiental de se treinar e utilizar grandes modelos como Codex-12B (GPT-afinado para geração de código com 12 bilhões de parâmetros). Por exemplo, estima-se que o treinamento de GPT-3 produziu cerca de 552 toneladas métricas de dióxido de carbono, o equivalente a o que mais de 120 carros produzem em um ano;[30]
- Os autores avaliaram qual a probabilidade de o modelo treinado gerar código idêntico a código encontrado em repositórios públicos

[30] Contudo, por mais que o treinamento de grandes modelos como GPT-3 necessitem de grandes quantidades de energia, sua inferência em, por exemplo, gerar 100 páginas de conteúdo, pode custar na ordem de 0,4 kW/h.

(GitHub). Algo que poderia vir a gerar implicações legais indesejadas (i.e., violação de direitos de propriedade privada).

Dados todos os possíveis riscos documentados, os autores ainda afirmam que:

> [...] dado o que foi exposto, modelos como o Codex devem ser desenvolvidos, utilizados, e suas capacidades exploradas cuidadosamente com o objetivo de maximizar seus impactos sociais positivos e minimizando os danos intencionais ou não intencionais que seu uso pode causar. Uma abordagem contextual é fundamental para análise de risco e mitigação eficazes, embora algumas categorias de mitigações sejam importantes a considerar em qualquer implantação de modelos de geração de código (Chen et al., 2021, p. 13).

Esse é um bom exemplo de um produto que foi desenvolvido sobre um regime de ética e segurança robusto. Robusto no sentido de que os problemas e limitações do modelo criado são (dentro do possível) conhecidos pelos desenvolvedores, que por sua parte tomaram a iniciativa de reportá-los à comunidade interessada. [31]

Esse é um dos papéis do engenheiro de segurança em IA. Não apenas avaliar os possíveis vieses e problemas que podem surgir durante o treinamento de um modelo e após sua implementação em um determinado contexto, *mas buscar mitigar novos problemas que possam vir a surgir.*

Nem sempre todos os potenciais usos de um modelo são de conhecimento de seus desenvolvedores. Talvez os primeiros modelos de aprendizagem

[31] Como outro exemplo, podemos citar a Redwood Research, uma organização que realiza pesquisa aplicada de alinhamento em IA. Em 2021, a organização estava desenvolvendo técnicas para controlar modelos geradores de texto (e.g., GPT-3), de modo a evitar que tais modelos produzam textos com conteúdo indesejado (o objetivo do modelo sendo treinado pelo projeto fora detectar quando um texto continha algum tipo de violência). Mais informações em: https://www.alignmentforum.org/posts/k7oxdbNaGATZbtEg3/redwood-research-s-current-project.

de máquina tivessem limites de utilização claros (e.g., classificar imagens de dígitos). Contudo, o mesmo não é verdade para os modelos sendo gerados hoje em dia. Muitas vezes modelos são capazes de realizar tarefas muito além daquelas que seus desenvolvedores tinham em mente. Citando mais uma vez o modelo treinado pela OpenAI, GPT-3 foi apenas treinado para "prever a próxima palavra de uma sequência". Esperava-se que o modelo seria proficiente em tarefas ligadas a NLP. O que não se esperava era que o modelo tivesse "aprendido" aritmética sem supervisão explícita.

Para evitar sermos pegos de forma desprevenida, análises de segurança devem ir um pouco além das ferramentas translacionais que revisamos. Precisamos de métodos quantitativos para avaliar, estressar e atacar nossos modelos. Mas como podemos implementar este tipo de prática no desenvolvimento de sistemas inteligentes? Na próxima seção, veremos uma ferramenta para esta tarefa.

Relatórios de Segurança e Cartas de Modelo

Dado que, em certos contextos e aplicações, a utilização de sistemas de IA deve ser monitorada de forma robusta. Uma forma de conectar as preocupações e apontamentos de desenvolvedores com aqueles que utilizarão tais modelos e aplicações, envolve a criação de uma documentação que detalhe as características de desempenho de um determinado sistema de IA, i.e., cartas de modelo.

Podemos definir uma carta de modelo como:

> [...] pequenos documentos acompanhando modelos treinados por aprendizagem de máquina que fornecem uma avaliação comparativa em uma variedade de condições, tais como entre diferentes grupos culturais, demográficos ou fenotípicos (e.g., raça, localização geográfica, sexo, tipo de pele) e grupos interseccionais (e.g., idade, sexo) que sejam relevantes para os domínios de aplicação pretendidos. Cartas de modelo também revelam o contexto no qual os modelos são destinados a serem utilizados, detalhes dos procedimentos de avaliação de desempenho, e outras informações relevantes (Mitchell et al., 2019, p. 220).

Podemos pensar em uma carta de modelo como o resultado final de uma avaliação de segurança de um determinado sistema de IA. Por mais que ainda não existam modelos de documentação padronizados e universais, existem sugestões de como tais modelos devem ser, e que tipo de informação deve estar explícita em uma carta modelo (Bender & Friedman, 2018; Holland et al., 2018; Gebru et al., 2018).

O intuito de uma carta de modelo é fornecer aos usuários de um determinado sistema informações sobre:

- Como utilizar o modelo;
- Como *não* utilizar o modelo;
- Os tipos de erros que o modelo pode cometer com mais frequência (i.e., suas vulnerabilidades).

Informados dessa realidade, espera-se que usuários sejam capazes de utilizar *"modelos imperfeitos"* da melhor forma possível. Cartas de modelos também podem beneficiar diversos tipos diferentes de atores:

- *Desenvolvedores de IA* podem entender melhor o quão bem um modelo pode funcionar para uma aplicação pretendida, comparar os resultados do modelo com outros modelos similares, entender como um modelo pode ser melhorado, afinado, e combinado com outros modelos;
- *Desenvolvedores de software* que utilizam das previsões de sistemas de IA podem melhor projetar suas aplicações;
- *Entidades reguladoras* podem entender como um sistema de IA pode falhar e impactar as pessoas, e utilizar tal informação para regulamentar o uso de IA para certas aplicações de alto risco;
- *Pessoas impactadas* por um sistema de IA podem utilizar de uma carta de modelo para determinar se os impactos experienciados foram devidamente previstos e especificados, e, ao mesmo tempo, saber quem são os responsáveis pelo desenvolvimento de tal modelo/aplicação.

Iremos nos basear no trabalho de Mitchell et al. (2019), *"Model Cards for Model Reporting"*, para demonstrar como tal ferramenta pode ser utilizada. No trabalho dos autores, Mitchell et al. (2019) utilizaram dois exemplos, um classificador de imagens (i.e., um detector de sorrisos) treinado com o conjunto de dados CelebA, e um modelo de detecção de toxicidade (e.g., detecção autônoma de textos com conteúdo tóxico).

Carta de Modelo
Detalhes do Modelo (informações básicas sobre o modelo)
1. Organização/Indivíduo que desenvolveu o modelo; 2. Data de desenvolvimento; 3. Versão do modelo (e.g., v 0.1); 4. Tipo de modelo (e.g., modelo de regressão logística, rede neural convolucional, modelo de linguagem transformer, vision transformer); 5. Informações sobre algoritmos de treinamento, parâmetros, características utilizadas, restrições de equidade ou outras abordagens aplicadas; 6. Artigo/Página do desenvolvedor/repositório do GitHub; 7. Informações para citação; 8. Licença; 9. Onde enviar perguntas e comentários sobre o modelo.
Uso Pretendido (casos de uso que foram previstos durante o desenvolvimento)
1. Uso pretendido primário (Qual o uso pretendido deste modelo?); 2. Usuários primários pretendidos (Qual o público alvo pretendido deste modelo?); 3. Utilizações fora da distribuição pretendida (Quais os tipos de aplicação que o modelo não foi treinado para dar suporte?).
Fatores (e.g., grupos demográficos, fenótipos, condições ambientais, atribuições técnicas ou outros fatores relevantes)
1. Fatores relevantes (Quais são os fatores para os quais o desempenho do modelo pode variar, e como estes foram determinados?); 2. Fatores de avaliação (Quais fatores estão sendo relatados, e por que estes foram escolhidos?).
Métricas (métricas devem ser escolhidas para refletir os impactos potenciais do modelo no mundo real)
1. Performance do modelo (e.g., acurácia, precisão, recall, AUC, etc.); 2. Limiares de decisão (Se forem utilizados limiares de decisão, quais são eles, e por que foram escolhidos?); 3. Abordagens de variação (Como a variabilidade do modelo foi medida? Desvio padrão? Variância?).
Dados de avaliação (detalhes do conjunto de dados utilizado para o treinamento e avaliação do modelo)
1. Conjunto de dados (Que conjunto de dados foi utilizado para avaliar o modelo?);

2. Motivação (Por que tal conjunto de dados foi escolhido?);
3. Pré-processamento (Como os dados foram pré-processados? Tokenização? Normalização? Amostras com valores "NaN" foram excluídas, ou seus valores foram estimados?);
4. Dados de treinamento (Nem sempre é possível fornecer tal conjunto. Quando possível, esta seção deve refletir os dados de avaliação. Se tal detalhe não for possível, informações mínimas permitidas devem ser fornecidas aqui, tais como detalhes da distribuição por vários fatores (e.g., distribuição de subgrupos entre características).

Considerações Éticas (uma análise ética não precisa necessariamente produzir soluções precisas, mas o processo de contemplação ética deve ser voltado para informar partes interessadas sobre preocupações levantadas pelos desenvolvedores e passos para trabalhos futuros)

1. O modelo utiliza algum dado sensível?
2. O modelo pretende informar decisões sobre questões centrais para a vida humana?
3. Que estratégias de mitigação de riscos foram utilizadas durante o desenvolvimento do modelo?
4. Que riscos podem estar presentes no uso do modelo?

Detalhes e Recomendações (preocupações adicionais que não foram cobertas nas seções anteriores)

1. Os resultados sugerem mais algum teste?
2. Havia algum grupo relevante que não foi representado no conjunto de dados de avaliação?
3. Existem recomendações adicionais para o uso do modelo?

Análise Quantitativa (análises quantitativas devem fornecer os resultados da avaliação do modelo de acordo com as métricas escolhidas, discriminadas pelos fatores escolhidos)

1. Resultados unitários (Como o modelo desempenhou com respeito a cada fator?);
2. Resultados interseccionais (Como o modelo desempenhou com respeito à interseção dos fatores avaliados?).

Na carta acima (Mitchell et al., 2019, p. 222), vemos diversos tipos de informações que podem esclarecer questões sobre o desenvolvimento, uso pretendido e possíveis problemas de um determinado modelo. Contudo, é importante lembrar que a lista acima não é exaustiva ou completa, e que tais relatórios devem ser sensíveis a um contexto de desenvolvimento/aplicação.

Por exemplo, a quantidade de informação que uma empresa privada está disposta a tornar público (e.g., dados de treinamento) pode ser menor do que uma organização acadêmica. Certas empresas podem vir a escolher não revelar certas informações chave para o desenvolvimento de uma aplicação comercial (e.g., algoritmos de treinamento). Mesmo assim, existem formas de apresentar informação pertinente (e.g., a performance de um modelo), sem revelar informação de caráter confidencial (e.g., como tal modelo foi desenvolvido).

Neste trabalho, iremos utilizar dois exemplos diferentes:

- Um modelo para *aprovação de cartão de crédito*, e;
- Um modelo de *previsão de renda anual*.

Através destes exemplos, iremos sugerir algumas metodologias e ferramentas para se: (1) inspecionar um modelo treinado por aprendizagem de máquina; e (2) "preencher" uma carta de modelo. Contudo, é importante lembrar que (de forma alguma) as ferramentas e metodologias apresentadas nesses exemplos são a totalidade da Segurança da IA. Todavia, certamente que elas podem auxiliar desenvolvedores a implementar uma avaliação de segurança inicial.

Exemplo 1: Aprovação de Cartão de Crédito

Avaliação de aplicações para cartões de crédito é uma tarefa que bancos comerciais comumente utilizam de inteligência artificial para automatizar. Neste exemplo, iremos desenvolver um modelo de regressão logística (uma das técnicas mais comuns em aprendizagem de máquina) para resolver um problema de classificação binária: classificar uma solicitação de cartão de crédito (caracterizada com uma série de características/features) como "Aprovada" ou "Reprovada".

Iremos utilizar o "Credit Approval Data Set" do Repositório de Aprendizagem de Máquina da UCI.[32] Este conjunto de dados possui 689 amostras de aplicações para cartões de crédito, rotuladas como aprovadas ou reprovadas. Contudo, para proteger a privacidade dos indivíduos que formam tal conjunto, todas as características foram mascaradas, i.e., ao invés de utilizarmos rótulos de características explícitas (e.g., `Gênero = ['Masculino', 'Feminino', 'Não-Binário']`), tais valores foram substituídos por símbolos (e.g., `Gênero = ['a', 'b', 'ab']`).

As próprias características/features foram removidas. Contudo, para este exemplo, trataremos cada amostra como formada pelas seguintes características (tipicamente solicitadas em aplicações para cartões de crédito):

- `"Gênero", "Idade", "Dívida", "Casado", "Cliente do Banco", "Nível de Educação", "Raça", "Anos de Emprego", "Inadimplência Prévia", "Empregado", "Crédito",`

[32] UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems. Disponível em: http://archive.ics.uci.edu/ml/datasets/credit+approval.

"Carteira de Motorista", "Cidadão", "Código Postal", "Renda";

E como alvo:

- "Status de Aprovação".

Os dados podem ser inicialmente visualizados como um Pandas[33] data frame:

	Gênero	Idade	Dívida	Casado	Cliente do Banco	Nível de Educação	Raça	Anos de Emprego	Inadimplência Prévia	Empregado	Crédito	Carteira de Motorista	Cidadão	Código Postal	Renda	Status de Aprovação
0	a	58.67	4.460	u	g	q	h	3.04	t	t	6	f	g	00043	560	+
1	a	24.50	0.500	u	g	q	h	1.50	t	f	0	f	g	00280	824	+
2	b	27.83	1.540	u	g	w	v	3.75	t	t	5	t	g	00100	3	+
3	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	00120	0	+
4	b	32.08	4.000	u	g	m	v	2.50	t	f	0	t	g	00360	0	+
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
684	b	21.08	10.085	y	p	e	h	1.25	f	f	0	f	g	00260	0	-
685	a	22.67	0.750	u	g	c	v	2.00	f	t	2	t	g	00200	394	-
686	a	25.25	13.500	y	p	ff	ff	2.00	f	t	1	t	g	00200	1	-
687	b	17.92	0.205	u	g	aa	v	0.04	f	f	0	f	g	00280	750	-
688	b	35.00	3.375	u	g	c	h	8.29	f	f	0	t	g	00000	0	-

689 rows × 16 columns

Como foi dito, características (especialmente aquelas categóricas) foram mascaradas por "símbolos sem sentido". Funções como `.info( )` e `describe( )` podem nos prover uma visão mais detalhada sobre os tipos de dados que estaremos trabalhando.

Rapidamente podemos ter acesso a quantos subgrupos cada característica possui (e.g., `Gênero = 3`, `Nível de Educação = 15`, `Raça = 10`, `Inadimplência Prévia = 2`) e outros dados estatísticos importantes (e.g., média, desvio padrão, valores máximos, valores mínimos).

[33] Pandas é uma biblioteca de Python para análise de dados.

	Gênero	Idade	Dívida	Casado	Cliente do Banco	Nível de Educação	Raça	Anos de Emprego	Inadimplência Prévia	Empregado	Crédito	Carteira de Motorista	Cidadão	Código Postal	Renda	Status de Aprovação
count	689	689	689.000000	689	689	689	689	689.000000	689	689	689.000000	689	689	689	689.000000	689
unique	3	349	NaN	4	4	15	10	NaN	2	2	NaN	2	3	170	NaN	2
top	b	?	NaN	u	g	c	v	NaN	t	f	NaN	f	g	00000	NaN	-
freq	467	12	NaN	518	518	137	398	NaN	360	395	NaN	373	624	132	NaN	383
mean	NaN	NaN	4.765631	NaN	NaN	NaN	NaN	2.224819	NaN	NaN	2.402032	NaN	NaN	NaN	1018.862119	NaN
std	NaN	NaN	4.978470	NaN	NaN	NaN	NaN	3.348739	NaN	NaN	4.866180	NaN	NaN	NaN	5213.743149	NaN
min	NaN	NaN	0.000000	NaN	NaN	NaN	NaN	0.000000	NaN	NaN	0.000000	NaN	NaN	NaN	0.000000	NaN
25%	NaN	NaN	1.000000	NaN	NaN	NaN	NaN	0.165000	NaN	NaN	0.000000	NaN	NaN	NaN	0.000000	NaN
50%	NaN	NaN	2.750000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN	0.000000	NaN	NaN	NaN	5.000000	NaN
75%	NaN	NaN	7.250000	NaN	NaN	NaN	NaN	2.625000	NaN	NaN	3.000000	NaN	NaN	NaN	396.000000	NaN
max	NaN	NaN	28.000000	NaN	NaN	NaN	NaN	28.500000	NaN	NaN	67.000000	NaN	NaN	NaN	100000.000000	NaN

Também é importante sabermos sobre o tipo de dados/características que estaremos trabalhando. Neste exemplo, estamos lidando com valores numéricos (números Inteiros, i.e., `int64`, números Reais, i.e., `float64`) e valores categóricos (classes, i.e., `object`).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 689 entries, 0 to 688
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Gênero                689 non-null    object
 1   Idade                 689 non-null    object
 2   Dívida                689 non-null    float64
 3   Casado                689 non-null    object
 4   Cliente do Banco      689 non-null    object
 5   Nível de Educação     689 non-null    object
 6   Raça                  689 non-null    object
 7   Anos de Emprego       689 non-null    float64
 8   Inadimplência Prévia  689 non-null    object
 9   Empregado             689 non-null    object
 10  Crédito               689 non-null    int64
 11  Carteira de Motorista 689 non-null    object
 12  Cidadão               689 non-null    object
 13  Código Postal         689 non-null    object
 14  Renda                 689 non-null    int64
 15  Status de Aprovação   689 non-null    object
dtypes: float64(2), int64(2), object(12)
memory usage: 86.2+ KB
```

O conjunto de dados utilizado neste exemplo possui uma série de valores faltando (exatamente 67) que podem prejudicar a performance do nosso modelo. Uma "boa prática" em ciência de dados e aprendizagem de máquina é: (1) remover as amostras com valores faltando; ou (2) substituir os valores faltando pôr os valores médios (e.g., o valor/categoria mais frequente) de cada característica.

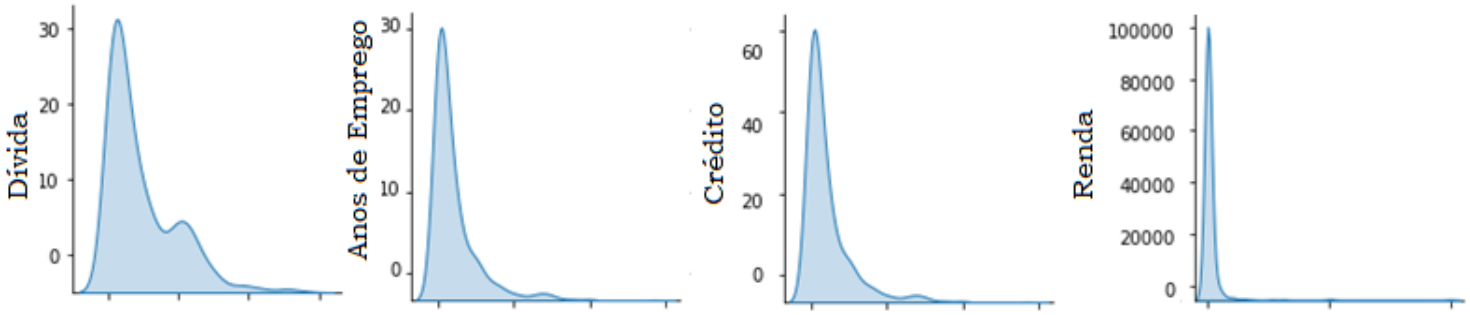
Como estamos trabalhando com um conjunto de dados pequeno, iremos utilizar a prática 2. Isto é um dos processos que fazemos durante pré-processamento, além de tornar todas as características categóricas em características numéricas.[34] Após essa fase, obtemos um conjunto de dados (com nenhum valor faltando) pronto para ser utilizado para treinarmos um modelo de classificação probabilística.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 689 entries, 0 to 688
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Gênero                 689 non-null    int64
 1   Idade                  689 non-null    int64
 2   Dívida                 689 non-null    float64
 3   Casado                 689 non-null    int64
 4   Cliente do Banco       689 non-null    int64
 5   Nível de Educação      689 non-null    int64
 6   Raça                   689 non-null    int64
 7   Anos de Emprego        689 non-null    float64
 8   Inadimplência Prévia   689 non-null    int64
 9   Empregado              689 non-null    int64
 10  Crédito                689 non-null    int64
 11  Carteira de Motorista  689 non-null    int64
 12  Cidadão                689 non-null    int64
 13  Código Postal          689 non-null    int64
 14  Renda                  689 non-null    int64
 15  Status de Aprovação    689 non-null    int64
dtypes: float64(2), int64(14)
memory usage: 86.2 KB
```

Podemos utilizar outras ferramentas para explorar os dados que estaremos trabalhando. Por exemplo, a biblioteca de visualização de

[34] Modelos de regressão logística não processão variáveis categóricas que não estejam codificadas como números.

dados Seaborn pode ser utilizada para explorar a distribuição dos dados que utilizaremos para treinar e avaliar nosso modelo.



Muitas das distribuições que temos possuem “longas caudas”, i.e., a distribuição de valores/amostras obedece a uma distribuição de Pareto, ou seja, o volume de amostras diminui conforme os valores aumentam. Disso, podemos interpretar que, por exemplo, a grande maioria das amostras: (i) não trabalhou por muitos anos; (ii) tem um score de crédito baixo; (iii) possui uma renda pequena (ou não declarada).

Em outras palavras, nosso conjunto de dados não é “uniforme”. Ele é extremamente enviesado, sendo um reflexo de um ambiente desigual (e.g., vieses históricos), e isto é uma bandeira vermelha. Nosso modelo pode vir a operar de forma menos eficiente quando lidando com amostras que não foram “vistas o suficiente” em seu treinamento (i.e., “statistical outliers”). Com isso em mente, uma análise mais profunda dos dados que estaremos trabalhando se mostra necessária.

Outra ferramenta que podemos utilizar para explorar os dados que estamos utilizando é a biblioteca *Facets*.

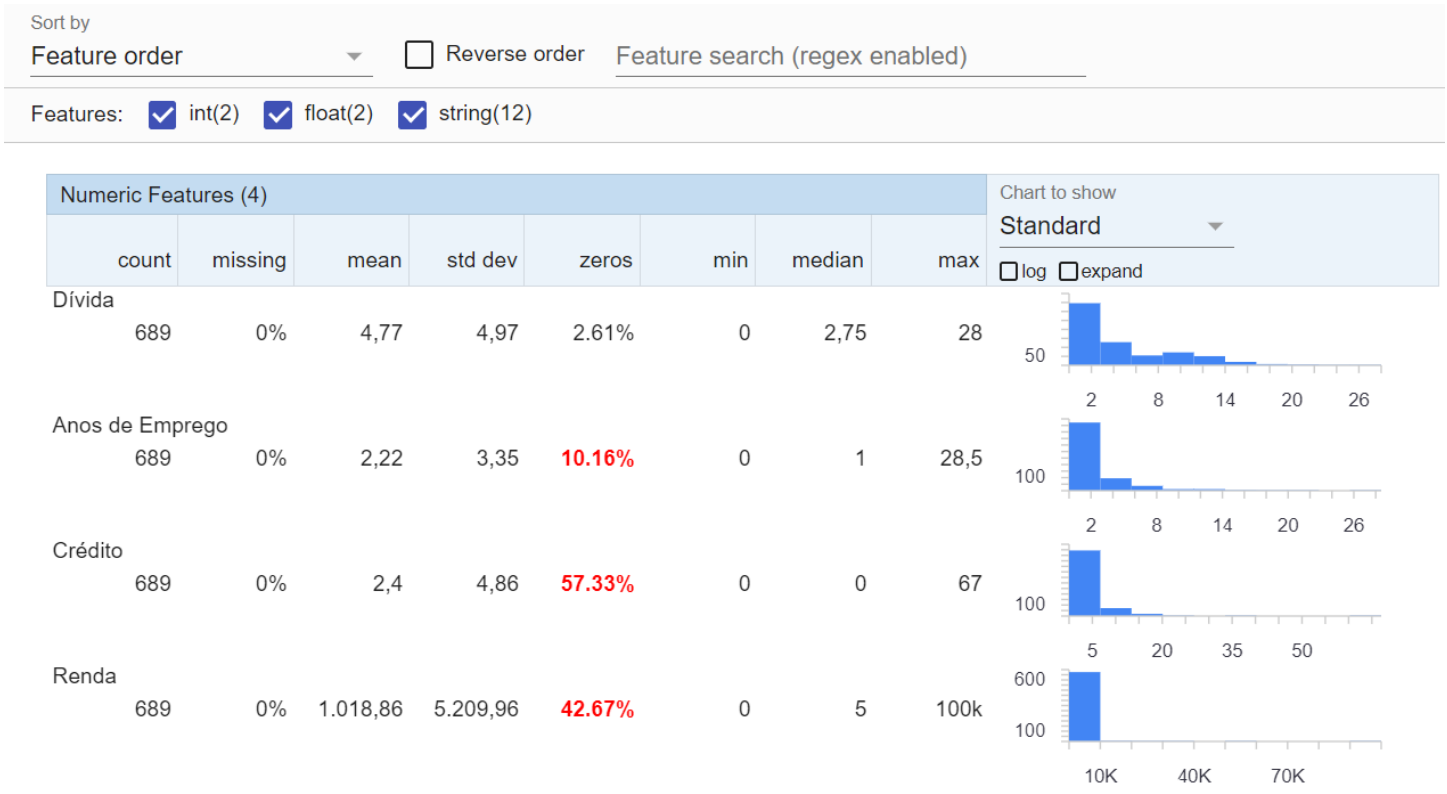
Facets[35] é uma ferramenta de código aberto para visualização de dados criada pela PAIR, desenvolvida para auxiliar na compreensão e análise de

[35] Disponível em: https://Github.com/PAIR-code/facets, ou em https://pair-code.Github.io/facets/.

conjuntos de dados utilizados em aprendizagem de máquina. Facets contém duas ferramentas de visualização:

- *Facets Overview:* Overview oferece uma forma rápida de explorar a distribuição de valores entre características de um conjunto de dados (e.g., valores comuns/incomuns, valores inesperados/ausentes, enviesamento);
- *Facets Dive:* Dive oferece uma interface interativa para explorar a relação entre características diferentes (e.g., como é a distribuição de Status de Aprovação versus Gênero e Raça?) e até mesmo amostras individuais.

Vejamos como as distribuições analisadas pela biblioteca Seaborn são apresentadas pela Facets Overview:



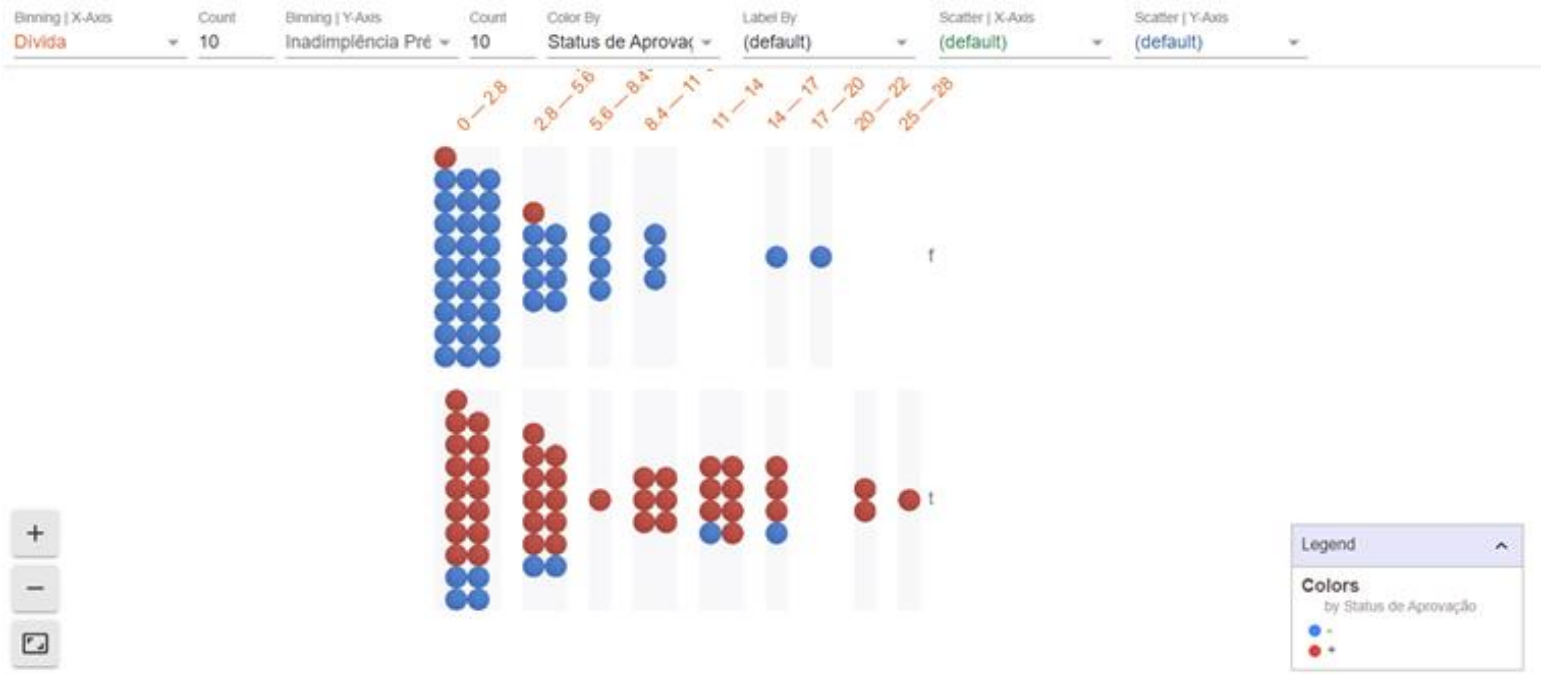
Novamente, longas caudas e distribuições enviesadas. Também é curioso que 42.67% das amostras tem “0” de renda. Se quase metade das amostras possui um valor nulo, deveríamos utilizar tal característica para treinar nosso classificador? Outra bandeira vermelha.

Ao mesmo tempo, mais da metade das amostras são de um gênero específico (“b”). Enquanto isso, das 689 amostras, 383 (55,5%) das solicitações de cartão de crédito foram negadas, e 306 (44,5%) solicitações foram aprovadas. Com este conjunto de dados, podemos estar criando um modelo que: (1) não possui um desempenho igual (e.g., paridade preditiva) entre gêneros; e (2) possui uma tendência maior por reprovar solicitações (Como isso pode afetar os clientes do banco?).

Agora, mergulhamos um pouco mais a fundo em nosso conjunto de dados com o Facets Dive. Como a característica “Raça” se relaciona com nosso alvo (Status de Aprovação)?

Alguns subgrupos da característica “Raça” estão fortemente sub representados (praticamente todas quando comparadas com o subgrupo “v”). Enquanto isso, alguns subgrupos apenas possuem exemplos negativos (Reprovados) enquanto outros apenas possuem exemplos positivos (Aprovados).

”Inadimplência prévia”, i.e., se o cliente deixou de pagar contas de outros cartões de crédito, deveria ser um fator determinante para uma solicitação de cartão de crédito, assim como “Dívida”. Como ambas essas características se relacionam com Status de Aprovação?



Tais características, aparentemente, são os fatores decisivos para inferirmos o Status de Aprovação de uma amostra, já que praticamente todas as amostras, conforme sua dívida aumenta (0-22), se encontram quase que totalmente divididas entre amostras que possuem inadimplência prévia (quase todos recebem um Status de Aprovação negativo) e não possuem inadimplência prévia (Status de Aprovação maioritariamente positivo).

Técnicas de visualização de dados podem nos trazer valiosos insights sobre o conjunto de dados que estamos trabalhando, seja para detectar possíveis falhas que nosso modelo pode vir a ter, ou decidindo quais as melhores características para usarmos em nosso modelo. Por exemplo, se adotarmos uma visão de justiça como "véu da ignorância", podemos escolher não utilizar nenhum atributo sensível para treinarmos nosso modelo (e.g., gênero, raça), já que aparentemente, "Inadimplência prévia" e "Dívida" possuem uma forte correlação com Status de Aprovação.

Por simplicidade, iremos treinar um modelo de regressão logística genérico utilizando scikit-learn, uma biblioteca de aprendizagem de máquina de código aberto. Os dados já foram pré-processados (e (re)escalados para valores pequenos, i.e., um número Real entre 0 e 1), e divididos entre um conjunto de treinamento (70% das amostras) e teste (30% das amostras). Nós não faremos validação neste exemplo, pois ele é apenas um "exemplo". Contudo, aplicações reais necessitam de fases de validação para afinação dos hiperparâmetros do modelo.

Para este exemplo, usaremos todas as 15 características disponibilizadas pelo conjunto de dados, pois será valioso para este estudo explorar como diferentes características se relacionam, e quais os coeficientes aprendidos pelo modelo para cada característica.

Coeficientes de correlação medem a associação linear entre variáveis/características. Podemos interpretar tais valores da seguinte forma:

- 1: Total correlação positiva;
- 0.8: Forte correlação positiva;
- 0.6: Correlação positiva moderada;
- 0: Sem qualquer correlação;
- -0.6: Correlação negativa moderada;
- -0.8: Forte correlação negativa;
- -1: Total correlação negativa.

Por exemplo, é ilegal definir o status de aprovação para uma solicitação de cartão de crédito com base na raça ou gênero do solicitante. Um valor positivo ou negativo do coeficiente de correlação dessas características com Status de Aprovação, significaria imparcialidade e descriminação por parte do banco que produziu este conjunto de dados (algo que não deve ser replicado por modelo nenhum). Por sorte, coeficientes de correlação podem ser facilmente calculados através da biblioteca NumPy[36] pela função `.corrcoef( )`.

Coeficientes de Correlação (Status de Aprovação)	
Gênero	0.0300
Idade	-0.1300
Dívida	-0.2000
Casado	0.1900
Cliente do Banco	0.1800
Nível de Educação	-0.1200
Raça	0.0003

[36] Uma biblioteca que fornece uma grande variedade de funções matemáticas (e.g., operações com matrizes multidimensionais) por comandos de alto nível.

Anos de Emprego	-0.3200
Inadimplência Prévia	-0.7100
Empregado	-0.4500
Crédito	-0.4000
Carteira de Motorista	-0.0300
Cidadão	0.1000
Código Postal	0.0900
Renda	-0.1700

Felizmente, aparentemente nenhum atributo sensível, como raça (0.0003) ou gênero (0.03), estão correlacionados com Status de Aprovação de forma significativa! Em contrapartida, a característica mais correlacionada com Status de Aprovação parece ser Inadimplência Prévia (-0.7100), algo que vai de acordo com nossa análise feita através da ferramenta Facets Dive. Aparentemente, os fatores determinantes para esse problema de classificação são “Inadimplência Prévia”, “Dívida”, “Empregado” e “Crédito”. Se determinarmos que tais atributos não são sensíveis, poderíamos muito bem treinar nosso classificador com apenas estas características, e ainda obter um resultado satisfatório.

Vejamos agora o resultado final do nosso modelo, i.e., sua performance com o conjunto de testes.

Performance (acurácia):		**0.85**
Matriz de Confusão	Classe prevista (Negativo)	Classe prevista (Positivo)
Classe verdadeira (Negativo)	**94**	**6**
Classe verdadeira (Positivo)	**26**	**102**

Alcançamos uma performance de 85%. Acima também vemos a matriz de confusão do teste que realizamos de nosso modelo. Já que treinamos nosso modelo com mais exemplos de "Reprovações" do que "Aprovações", podemos ver que nosso modelo possui uma tendência maior para classificar pessoas que deveriam ser aprovadas como reprovados (Falsos Negativos = 11%), do que aprovar pessoas que deveriam ser reprovadas (Falsos Positivos = 0.2%).

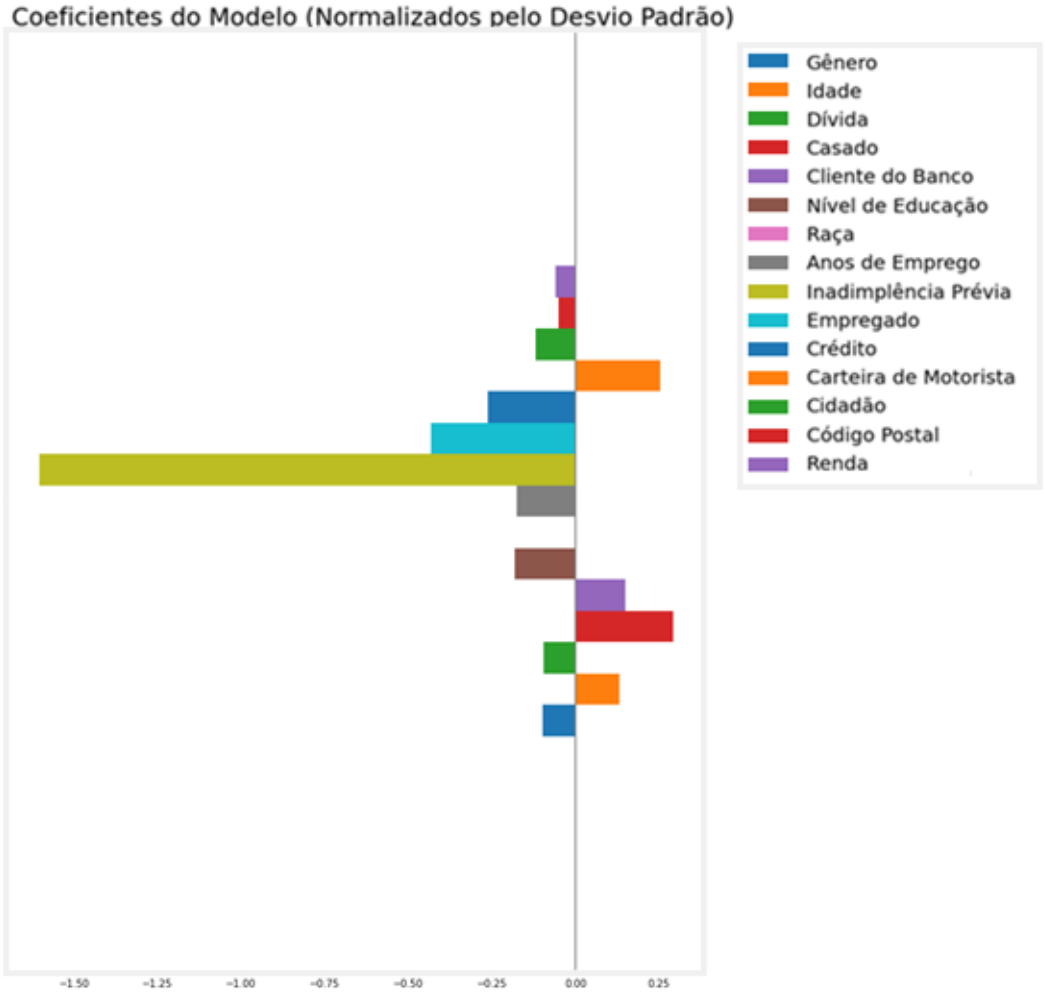
Talvez, uma sugestão para os gerentes de banco que utilizam desta ferramenta seja *"Reprovações devem ser melhor investigadas/analisadas, vocês podem estar perdendo um bom cliente"*. Contudo, se for do interesse do banco que Falsos Positivos sejam evitados ao máximo, o modelo treinado apresente uma boa relação entre verdadeiros positivos e falsos positivos (i.e., *Precisão = 0.94*).

Faremos apenas mais uma análise neste exemplo. Iremos analisar os coeficientes aprendidos pelo nosso modelo de regressão, que basicamente indicam, assim como os coeficientes de correlação, o quanto de "atenção" nosso modelo atribui a cada uma das características de uma amostra.

Contudo, antes de calcular tais coeficientes, precisamos normalizá-los. Para isso, utilizaremos de funções da biblioteca Pandas, `.var( )` e `.std()`, para calcular a variância e desvio padrão dos valores das nossas características. [37]

Desvio padrão e variância podem nos auxiliar a entender outras relações importantes do nosso conjunto de dados. E com o desvio padrão, podemos normalizar os coeficientes de nosso modelo e interpretá-los de forma correta (i.e., valores normalizados são valores que "partilham" de uma escala fictícia comum).

[37] Lembrando que a variância e desvio padrão foram calculadas com os valores reescalados/normalizados (delimitados entre 0 e 1), pois não haveria sentido comparar a variância e desvio padrão de valores medidos por escalas diferentes (e.g., anos versus dólares?).



Novamente, o principal fator para se prever "Status de Aprovação" é "Inadimplência Prévia". Veja que "Raça", com um coeficiente de -0.002, nem é visível na plotagem acima. Munidos de todas essas informações, vamos agora preencher nossa carta de modelo.

Carta de Modelo – Aprovação de Cartão de Crédito
Detalhes do Modelo
1. Modelo desenvolvido por Nicholas Kluge, pesquisador da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), em outubro de 2021;

2. Trata-se de um modelo de Regressão Logística para classificação binária, versão 0.1. Este modelo foi treinado para classificar solicitações de cartão de crédito como “Reprovadas” ou “Aprovadas”;
3. Este modelo foi treinado apenas por motivações acadêmicas, e ele não segue nenhum tipo de restrição de equidade/justiça. Este modelo não foi criado para ser implementado em aplicações reais;
4. O conjunto de dados utilizado é o Credit Approval Data Set da UCI Machine Learning Repository. Disponível em: http://archive.ics.uci.edu/ml/datasets/credit+approval;
5. O código para este modelo pode ser encontrado em: https://Github.com/Nkluge-correa/AI-Ethics-exercise;
6. Licença: MIT License;
7. Contato: nicholas.correa@acad.pucrs.br.

Uso Pretendido

1. O uso pretendido deste modelo, e o código compartilhado, é apresentar ao desenvolvedor ferramentas para se explorar um conjunto de dados, e avaliar possíveis implicações éticas e falhas de segurança de um modelo treinado por aprendizagem de máquina. Este modelo, e código, não foram criados para serem utilizados em aplicações reais. Contudo, as ferramentas utilizadas podem sim ser utilizadas para avaliações éticas de modelos treinados por aprendizagem de máquina;
2. Este modelo foi desenvolvido para o público acadêmico, desenvolvedores e praticantes de aprendizagem de máquina interessados em aprender como desenvolver modelos “justos”;
3. Como um experimento acadêmico, a única utilização para este modelo é a classificação de solicitações de cartão de crédito de amostras retiradas do Credit Approval Data Set Este modelo não deve ser usado para, e.g., classificação de score de crédito, inferência de score de crédito, ou qualquer outro tipo de tarefa diferente do seu uso primário pretendido.

Fatores

1. As características utilizadas para a tarefa de classificar o Status de Aprovação de um solicitante de cartão de crédito são: “Gênero”, “Idade”, “Dívida”, “Casado”, “Cliente do Banco”, “Nível de Educação”, “Raça”, “Anos de Emprego”, “Inadimplência Prévia”, “Empregado”, “Crédito”, “Carteira de Motorista”, “Cidadão”, “Código Postal”, “Renda”. Atributos como “Gênero” e “Raça” são considerados como atributos sensíveis;
2. Os dados utilizados para treinamento não possuem uma distribuição uniforme entre os subgrupos de cada característica. Existe um forte enviesamento, para certos tipos de subgrupos, como gêneros e raças específicas.

Métricas

1. A métrica de performance utilizada foi acurácia (nº total de classificações corretas por total de classificações realizadas), 85% de acerto durante a faze de teste;
2. O modelo possui uma tendência maior para classificar pessoas que deveriam ser aprovadas como reprovados (Falsos Negativos = 11%), do que aprovar pessoas que deveriam ser reprovadas (Falsos Positivos = 0.2%);
3. Sugestão: reprovações devem ser melhor investigadas/analisadas;
4. Dados de treinamento e testagem foram divididos do conjunto de dados fornecidos pela UCI Machine Learning Repository (i.e., Credit Approval Data Set);
5. Este conjunto de dados foi escolhido por sua disponibilidade pública.
6. Amostras com valores ausentes (i.e., "?" ou "NaN") tiveram tais valores substituídos pelo valor médio de sua característica específica.

Considerações Éticas

1. Dada a distribuição enviesada dos dados de treinamento, o modelo pode se comportar de forma ineficiente quando lidando com amostras pouco vistas;
2. O modelo utiliza de dados sensíveis (i.e., Raça e Gênero);
3. Recomenda-se que para aplicações reais, atributos sensíveis (e.g., raça e gênero) e atributos contendo valores "anormais" (e.g., renda) não sejam utilizados para classificação;
4. De acordo com os coeficientes de correlação, e coeficientes aprendidos pelo modelo, atributos sensíveis não interferem na classificação do modelo;
5. Os atributos mais correlacionados com o Status de Aprovação do solicitante são: "Inadimplência Prévia", "Dívida", "Empregado" e "Crédito".

Detalhes e Recomendações

1. Não foi realizada uma análise da performance do modelo entre diferentes subgrupos de cada característica. Uma análise mais aprofundada pode revelar que o modelo viola critérios de equidade, como, por exemplo, paridade preditiva;
2. Os dados utilizados para este exemplo não refletem o contexto social e histórico de um lugar como, por exemplo, Brasil. Eles refletem o contexto social e histórico Norte-Americano. Assim, não se recomenda utilizá-lo para desenvolvimento de aplicações fora deste domínio específico.

Análise Quantitativa

	Coeficientes
Gênero	-0.211754
Idade	0.476012
Dívida	-0.526039
Casado	1.753660
Cliente do Banco	0.510739
Nível de Educação	-0.568626
Raça	-0.002892
Anos de Emprego	-0.903885
Inadimplência Prévia	-3.210696
Empregado	-0.879579
Crédito	-1.566622
Carteira de Motorista	0.508445
Cidadão	-0.413766
Código Postal	-0.170551
Renda	-1.057520

	Coeficientes de Correlação
Gênero	0.0300
Idade	-0.1300
Dívida	-0.2000
Casado	0.1900
Cliente do Banco	0.1800
Nível de Educação	-0.1200
Raça	0.0003
Anos de Emprego	-0.3200
Inadimplência Prévia	-0.7100
Empregado	-0.4500
Crédito	-0.4000
Carteira de Motorista	-0.0300
Cidadão	0.1000
Código Postal	0.0900
Renda	-0.1700

Performance (acurácia) do modelo de regressão logística: 0.8596491228070176

	Classe prevista (Negativo)	Classe prevista (Positivo)
Classe verdadeira (Negativo)	94	6
Classe verdadeira (Positivo)	26	102

Exemplo 2: Previsão de Renda Anual

Algo que não fizemos na nossa última análise (Exemplo 1), foi avaliar/comparar a performance do modelo treinado entre diferentes subgrupos:

- *Gênero: como a performance do modelo difere entre homens e mulheres?*

Neste exemplo, faremos exatamente isto.

Iremos utilizar o "Adult Census Income Data Set",[38] também disponibilizado pela UCI Machine Learning Repository. Este conjunto de dados é um "clássico" da aprendizagem de máquina, extraído do Escritório do Censo dos EUA em 1994, por Ronny Kohavi e Barry Becker. A tarefa que iremos atacar também será uma tarefa de previsão binária: *determinar se uma pessoa ganha mais de $50.000 USD por ano.*

Iremos utilizar praticamente todas as bibliotecas que utilizamos no Exemplo 1 (i.e., Numpy, Pandas, Matplotlib, Seaborn, Facets), com a adição de duas bibliotecas novas: Tensorflow[39] e Keras.[40] As características contidas neste conjunto de dados são:

- "`age`" (Idade), "`workclass`" (classe trabalhadora), "`fnlwgt`" (o número de indivíduos que o Censo acredita que o conjunto de observações representa, i.e., o peso das observações), "`education`" (nível de escolaridade), "`education_num`" (uma

[38] Lichman, M. (2013). UCI Machine Learning Repository. Disponível em: http://archive.ics.uci.edu/ml/datasets/Census+Income.

[39] Uma biblioteca de código aberto para aprendizagem de máquina. Disponível em: https://www.tensorflow.org/.

[40] Uma biblioteca de código aberto, criada por François Chollet, para desenvolvimento de redes neurais. Disponível em: https://keras.io/.

enumeração da representação categórica da educação), "marital_status" (estado civil), "occupation" (ocupação), "relationship" (relacionamento com as pessoas da casa), "race" (raça), "gender" (gênero), "capital_gain" (ganhos de capital), "capital_loss" (perdas de capital), "hours_per_week" (horas trabalhadas por semana), "native_country" (nacionalidade), "income_bracket" (renda anual).

Temos 14 características e 1 alvo (i.e., renda anual).

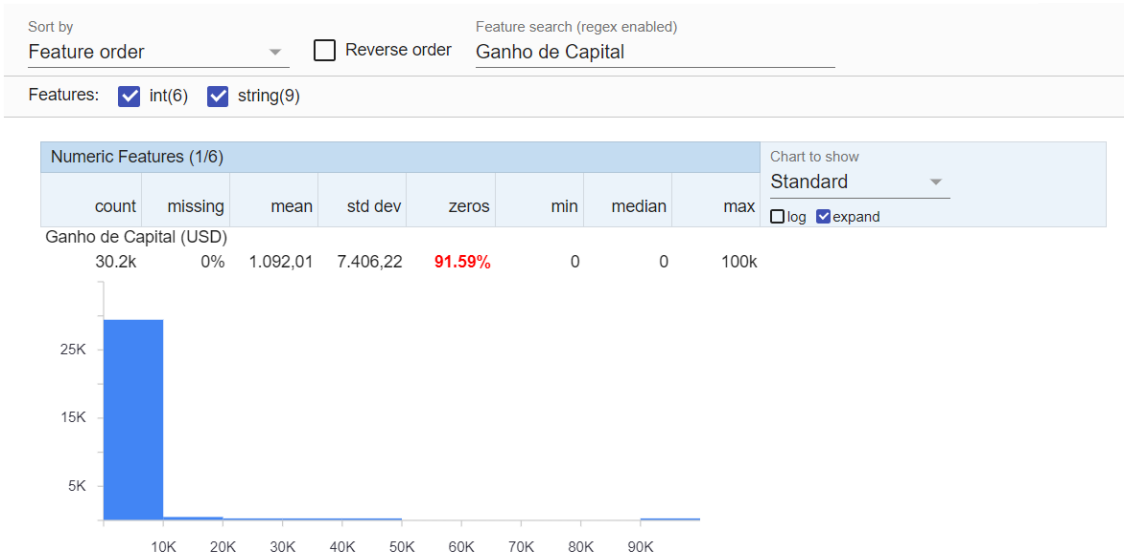
	Idade	Classe Trabalhadora	fnlwgt	Educação	Educação_num	Estado Civil	Ocupação	Parentesco	Raça	Gênero	Ganho de Capital (USD)	Perda de Capital (USD)	Trabalho (Horas/Semana)	Nacionalidade	Renda
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

Com este conjunto de dados, temos uma vantagem em relação ao conjunto utilizado no exemplo anterior: nós temos mais de 32 mil amostras para utilizar. Assim, dessa vez não iremos substituir valores incomuns/ausentes (e.g., 'NaN', '?') por seus respectivos valores médios, mas iremos excluir todas as amostras que possuam valores ausentes. Com isso, nos restam exatamente 30,163 amostras para treinamento (45,224 se contarmos com as amostras do conjunto de teste). E novamente, durante o pré-processamento, todas as características (categóricas) serão transformadas em números e normalizadas.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       30725 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education_num   32561 non-null  int64
 5   marital_status  32561 non-null  object
 6   occupation      30718 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   gender          32561 non-null  object
 10  capital_gain    32561 non-null  int64
 11  capital_loss    32561 non-null  int64
 12  hours_per_week  32561 non-null  int64
 13  native_country  31978 non-null  object
 14  income_bracket  32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

Antes disso, vamos inspecionar nosso conjunto de dados diretamente com o Facets, que é (de longe) a melhor ferramenta de análise e visualização de dados que apresentamos no exemplo anterior. Algumas perguntas que podem guiar nossa investigação são:

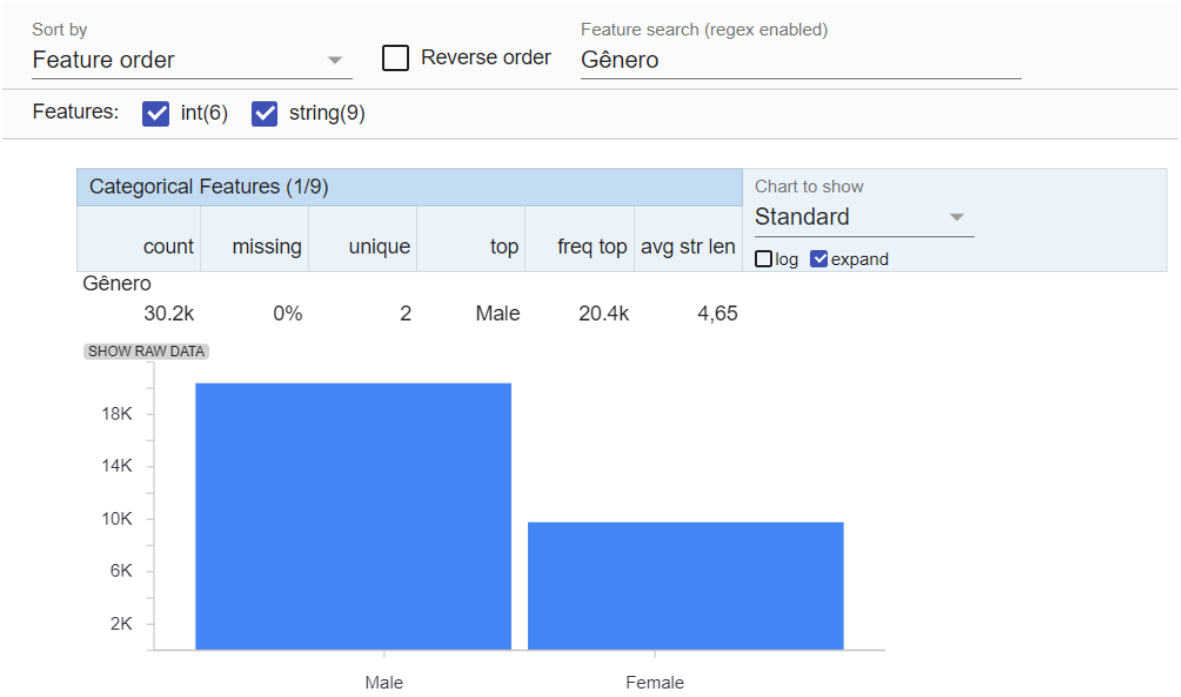
- *Há características em falta que podem afetar outras características?*



Definitivamente sim. Para ganho/perda de capital/investimentos, podemos ver que mais de 90% dos valores são 0. Em um mundo onde a distribuição de renda é extremamente desigual, não deve ser uma surpresa que menos de 10% tenham valores diferentes de 0. A grande maioria da população não investe, ganha ou perde capital (pois simplesmente não o possui). xxx

Contudo, não é nada óbvio como interpretar tal resultado. Afinal, “0” significa nenhum ganho/perda ou ganho/perda não declarado? Ambas as situações são bem diferentes. Em situações como esta, é melhor não utilizar tal característica para o treinamento do nosso modelo.

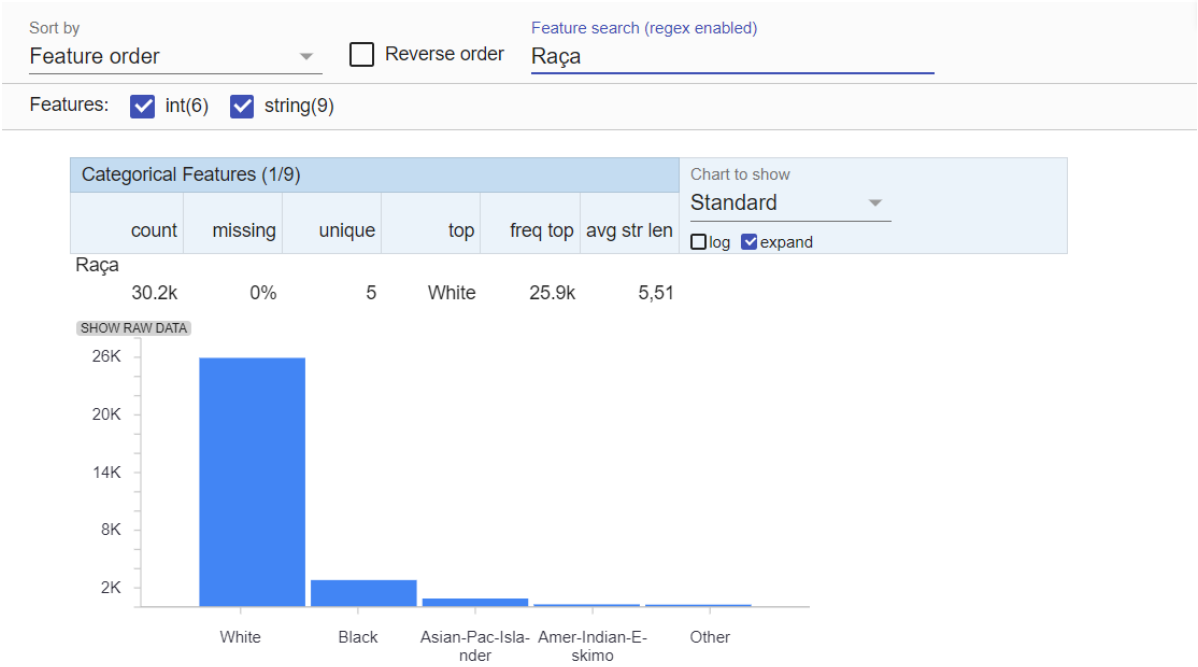
- *Existem sinais de enviesamento no conjunto de dados?*



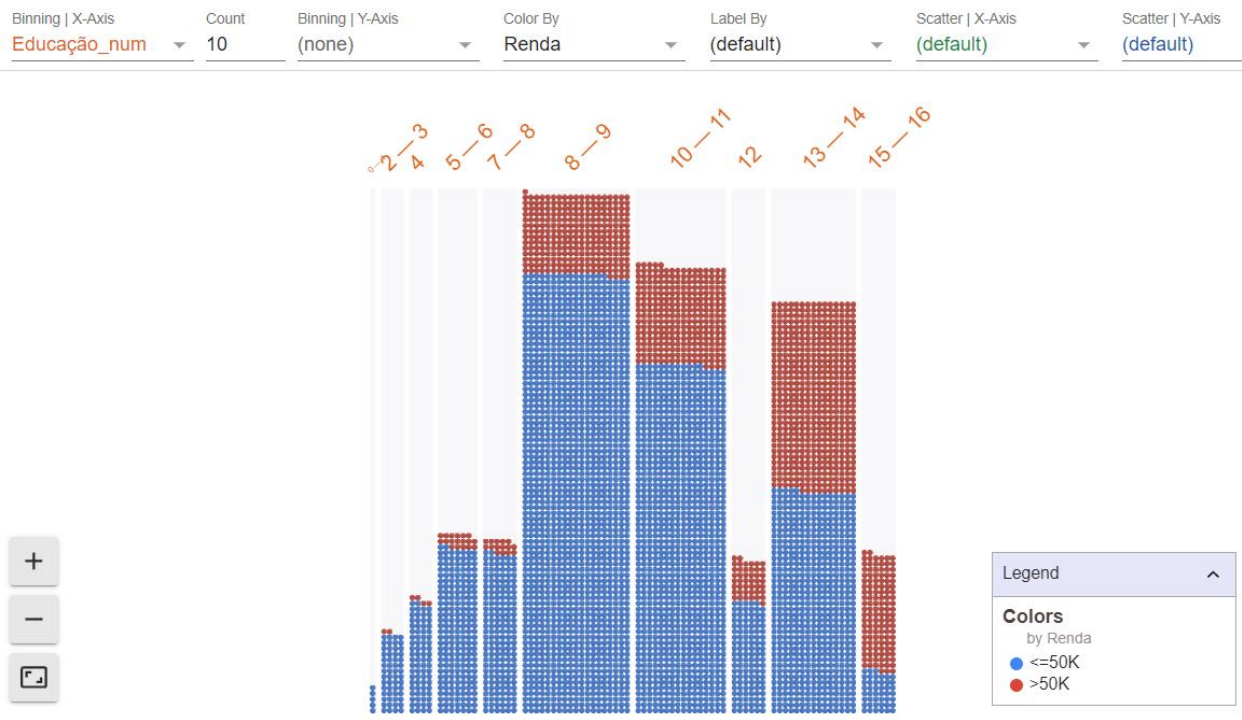
Sim. 67% dos exemplos representam homens. Isto sugere um considerável enviesamento nos dados, pois esperaríamos que a repartição entre os sexos fosse mais próxima de 1:1. Além da sub-

representação do gênero feminino, vemos uma grande sub-representação racial.

Esse enviesamento pode vir a prejudicar a performance do nosso modelo para com um subgrupo no qual existem poucas amostras/exemplos.

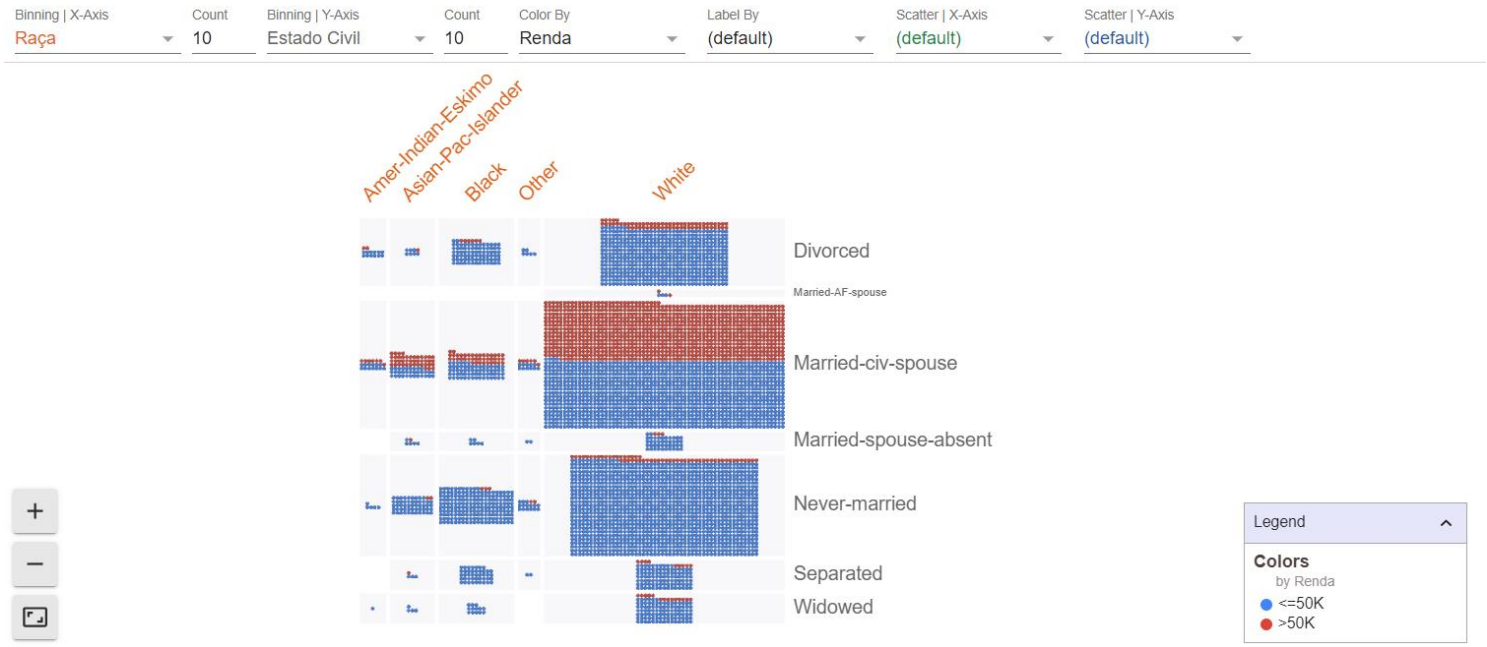


Utilizando o Facets Dive, podemos procurar por formas como as características estão correlacionadas umas com as outras. Renda anual e nível de escolaridade parecem estar bem correlacionados, já que para os maiores níveis de educação (e.g., doutorado e pós-doutorado), vemos a única classe onde a maioria das amostras recebe > $50,000 USD.



Enquanto isso, se explorarmos Ocupação × Gênero, veremos que raramente encontramos mulheres trabalhando no setor agrícola pecuário (Seria isto uma fiel representação do mundo real?), enquanto que mulheres dominam ocupações que envolvem cargos administrativos e clericais.

Existem muitas outras correlações para serem investigadas, uma última que mostraremos é a intersecção de amostras entre Raça × Estado Civil × Renda.



Em poucas palavras, se você quer encontrar as amostras com uma renda anual superior a $50,000 USD, procure por pessoas caucasianas casadas.

Para este exemplo, iremos apenas utilizar as seguintes características para treinar nosso modelo:

- `"workclass", "race", "education", "marital_status", "age", "relationship", "native_country", "occupation".`

E utilizaremos as bibliotecas Keras e TensorFlow para criar e treinar uma "densely connected feed forward neural network" (uma rede neural direta e densamente conectada) com três camadas ocultas (os parâmetros de afinação do modelo desenvolvido podem ser vistos no notebook deste exemplo). Utilizaremos 30,163 de amostras para treinamento e 15,061 amostras para testarmos o modelo (novamente, já que isto é apenas um exemplo, pularemos a fase de validação).

Neste exemplo, utilizaremos mais de uma métrica para avaliarmos a performance do nosso modelo: acurácia,[41] precisão,[42] recall[43] e AUC.[44] No geral, nosso modelo alcança os seguintes valores de performance:

	Acurácia	**Precisão**	**Recall**	**AUC**
Performance	0.8325	0.7074	0.5577	0.8832

Acurácia é a mesma métrica de performance que utilizamos no primeiro exemplo. Esta é a métrica mais "direta" e usualmente utilizada, *"quantas vezes o classificador acertou?"*. Contudo, nem sempre a acurácia é a métrica que devemos adotar para avaliar uma determinada aplicação.

Precisão é geralmente utilizada como métrica de performance para aplicações onde um falso positivo é um problema pior do que um falso negativo. Por exemplo, em detecção de spams um falso positivo significa bloquear um e-mail potencialmente importante. Enquanto que receber

[41] A fração de previsões que um modelo de classificação acertou. Em classificação binária, a acurácia tem a seguinte definição:

$$acc = \frac{Verdadeiros\ positivos + Verdadeiros\ Negativos}{número\ total\ de\ amostras}$$

[42] Precisão:

$$pre = \frac{Verdadeiros\ positivos}{Verdadeiros\ positivos + Falsos\ Positivos}$$

[43] Recall:

$$rec = \frac{Verdadeiros\ positivos}{Verdadeiros\ positivos + Falsos\ Negativos}$$

[44] AUC (Área sob a Curva ROC, i.e., uma curva da taxa de verdadeiros positivos versus a taxa de falsos positivos em diferentes limiares de classificação) é a probabilidade de um classificador estar mais confiante de que uma amostra positiva, escolhida aleatoriamente, é realmente positiva, do que uma amostra negativa, escolhida aleatoriamente, é positiva.

spam é algo “tolerável”, perder a aguardada resposta daquela revista acadêmica prestigiosa é inaceitável.

Recall é o oposto de precisão. Essa métrica mede falsos negativos contra verdadeiros positivos, e em aplicações como detecção de doenças, onde falsos negativos devem ser evitados a todo o custo, recall é a performance que devemos estar atentos.

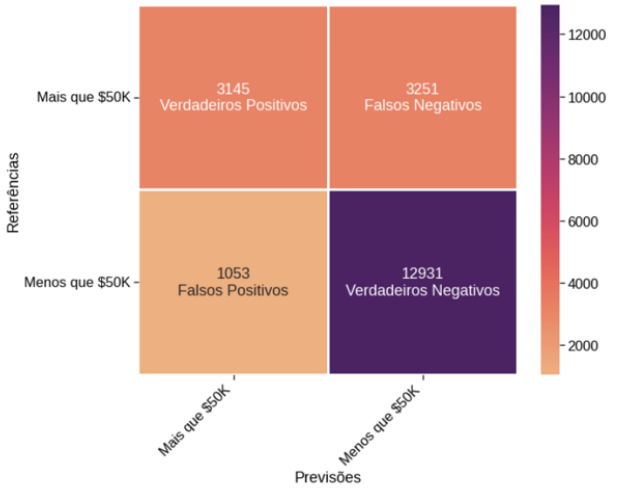
Já AUC, que no caso de nosso modelo é uma métrica com o valor mais próximo da acurácia, é a probabilidade de, digamos, nosso classificador gritar “Lobo!”, quando realmente há um lobo por perto. Ou seja, classificar uma amostra aleatoriamente selecionada como sua classe verdadeira.

Qual a métrica que devemos utilizar para avaliar nosso modelo? Depende da aplicação deste modelo. Digamos que o modelo será utilizado para avaliar quem (por ter uma renda anual > $50,000 USD) deve pagar mais impostos. Para essa aplicação, um falso positivo (o indivíduo é classificado como recebendo > $50,000 USD, mas na verdade, recebe < $50,000 USD) parece ser mais danoso do que um falso negativo. Ou seja, para essa aplicação, a precisão parece ser a métrica de performance adequada (por sorte, a precisão do nosso modelo é superior ao seu recall).

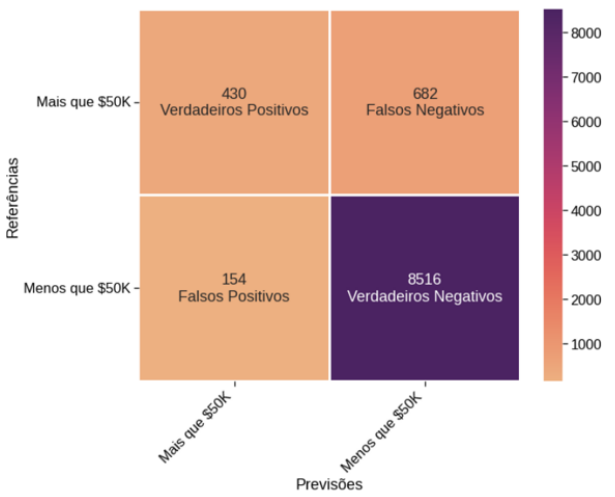
Embora a avaliação do desempenho geral do modelo nos dê alguma percepção de sua qualidade, ela não nos dá muita percepção do desempenho de nosso modelo para diferentes subgrupos. Avaliar uma rede neural profunda é diferente de avaliar um simples modelo de regressão logística, já que não podemos inspecionar os coeficientes deste modelo de forma inteligível e simples (nossa rede neural possui mais de 35 mil parâmetros treinados).

Neste exemplo, iremos definir que gênero, raça e estado civil são atributos sensíveis. E iremos explorar algumas das diferenças de performance entre subgrupos dessas características. Se formos comparar a matriz de confusão do subgrupo “Male” versus “Female”:

	Acurácia	Precisão	Recall	AUC
Male	0.7888	0.7492	0.4917	0.8592

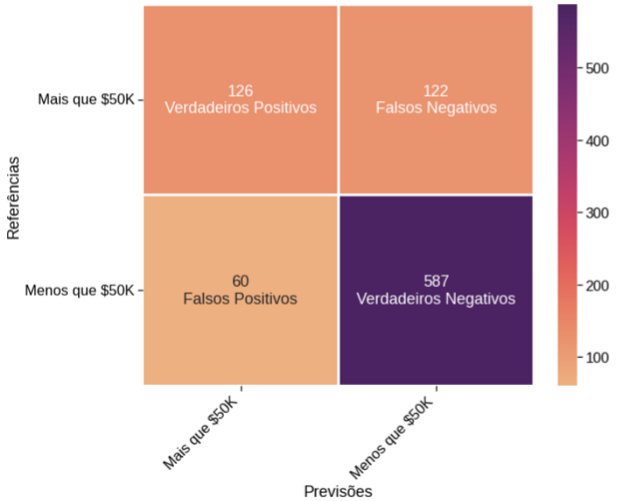


	Acurácia	Precisão	Recall	AUC
Female	0.9145	0.7363	0.3867	0.9118

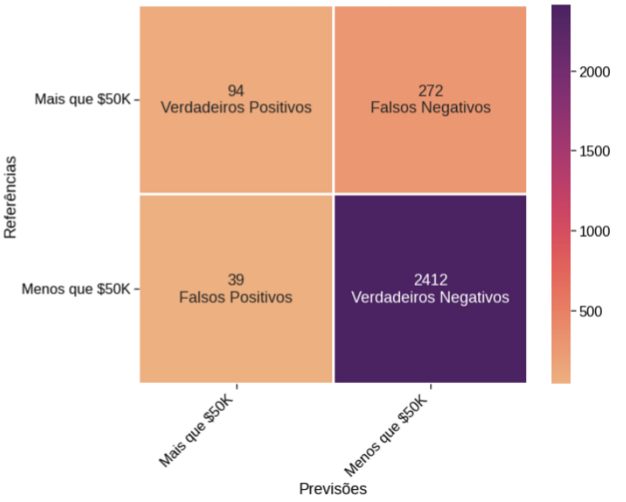


Veremos que, em termos de acurácia e AUC, mulheres recebem uma classificação melhor (precisão possuindo um valor quase igual para ambos os sexos). Contudo, como sabemos que mulheres são desproporcionalmente representadas neste conjunto de dados, isto é um possível sinal de sobreajuste. Também existe uma considerável discrepância na performance deste modelo entre subgrupos das características raça, gênero e estado civil.

	Acurácia	Precisão	Recall	AUC
Asian-Pac-Islander	0.7966	0.6774	0.5081	0.8592



	Acurácia	Precisão	Recall	AUC
Black	0.8896	0.7068	0.2568	0.9102



Contudo, um ponto positivo é que temos, no geral, um valor alto de precisão em combinação com um valor baixo de recall. Uma forma de interpretar este resultado é que nosso classificador é extremamente "exigente", no sentido de que todas as pessoas classificadas como "Renda anual > $50,000 USD", realmente possuem essa renda. No entanto, o modelo deixa de classificar positivamente diversas pessoas com renda > $50,000 USD, pois nosso modelo é "extremamente exigente".

Se utilizarmos este modelo para definirmos quem deve pagar mais (ou menos) impostos, quando o modelo classificar alguém como "Renda anual > $50,000 USD", o modelo quase sempre acertará (o modelo é preciso). Contudo, muitas pessoas que também possuem uma renda > $50,000 USD não serão "pegas" por este classificador.

O resumo da performance do modelo treinado, entre subgrupos dos atributos sensíveis determinados, é o seguinte:

Performance por Gênero				
	Acurácia	**Precisão**	**Recall**	**AUC**
Masculino	0.7888	0.7492	0.4917	0.8592
Feminino	0.9145	0.7363	0.3867	0.9120
Performance por Raça				
	Acurácia	**Precisão**	**Recall**	**AUC**
Caucasiano	0.8227	0.7527	0.4882	0.8812
Negro	0.8896	0.7068	0.2568	0.9102
Asiático-Americano	0.7966	0.6774	0.5081	0.8592
Esquimós	0.8951	0.6429	0.2647	0.7831
Outros	0.9134	0.5385	0.3333	0.9209

Performance por Estado Civil				
	Acurácia	**Precisão**	**Recall**	**AUC**
Casado (cônjuge civil)	0.7120	0.7475	0.5541	0.7900
Divorciado	0.8949	0.7143	0.0332	0.7959
Casado (cônjuge ausente)	0.9189	0.6667	0.0645	0.8214
Nunca Casado	0.9524	1.0000	0.0149	0.8859
Separado	0.9329	0.8000	0.0606	0.8442
Casado (cônjuge militar)	0.5238	0.0000	0.0000	0.6955
Viúvo	0.9033	0.5000	0.0125	0.7569

Não podemos atestar paridade estatística, paridade preditiva ou probabilidades equalizadas para este modelo. Os resultados demonstram que tal modelo não atende a estes critérios de equidade, já que, por exemplo, certos subgrupos são mais suscetíveis a certos erros de previsão do que outros (especialmente indivíduos que pertencem a certos subgrupos de estado civil, e.g., Casado-cônjuge-militar).

Tais resultados sugerem que temos um modelo que é sobre ajustado, muito em parte pela sub-representação de diversos subgrupos. Assim, não podemos garantir que tal modelo irá generalizar bem, pois não possuímos exemplos o suficiente de todos os subgrupos para que tal modelo “aprenda”.

Como todos estes resultados em mão, podemos agora preencher nossa carta de modelo:

Carta de Modelo – Predição de Renda Anual

Detalhes do Modelo

1. Modelo desenvolvido por Nicholas Kluge, pesquisador da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), em outubro de 2021;
2. Trata-se de uma rede neural profunda direta (densa), treinada para resolver uma tarefa de classificação binária, versão 0.1. Este modelo foi treinado para classificar indivíduos entre "Renda anual > $50,000 USD" ou "Renda anual < $50,000 USD";
3. Este modelo foi treinado apenas por motivações acadêmicas, e ele não segue nenhum tipo de restrição de equidade/justiça. Ele não foi criado para ser implementado em aplicações reais;
4. O conjunto de dados utilizado é o Adult Census Income Data Set, disponibilizado pela UCI Machine Learning Repository. Disponível em: http://archive.ics.uci.edu/ml/datasets/Census+Income;
5. O código para este modelo pode ser encontrado em: https://Github.com/Nkluge-correa/AI-Ethics-Exercice-2;
6. Licença: MIT License;
7. Contato: nicholas.correa@acad.pucrs.br.

Uso Pretendido

1. O uso pretendido deste modelo, e o código compartilhado, é apresentar ao desenvolvedor algumas ferramentas para se explorar um conjunto de dados, e avaliar possíveis implicações éticas e falhas de segurança de um modelo treinado por aprendizagem de máquina. Este modelo, e código, não foram criados para serem utilizados em aplicações reais. Contudo, as ferramentas utilizadas podem sim ser utilizadas para avaliações éticas de modelos treinados por aprendizagem de máquina;
2. Este modelo foi desenvolvido para o público acadêmico, desenvolvedores e praticantes de aprendizagem de máquina interessados em aprender como desenvolver modelos "justos";
3. Como um experimento acadêmico, a única utilização para este modelo é a predição de Renda Anual de amostras retiradas do Adult Census Income Data Set. Este modelo não deve ser usado para, e.g., predição de renda vitalícia, ou qualquer outro tipo de tarefa diferente do seu uso primário pretendido.

Fatores

1. As características utilizadas para o treinamento do modelo são: "Classe Trabalhadora", "Raça", "Educação", "Estado Civil", "Idade", "Parentesco", "Nacionalidade", "Ocupação". Atributos como "Gênero", "Raça" e "Estado Civil" foram considerados como atributos sensíveis;

2. Os dados utilizados para treinamento não possuem uma distribuição uniforme entre os subgrupos de cada característica. Existe um forte enviesamento, para certos tipos de subgrupos, como gêneros, estados civis e raças específicas.

Métricas

1. As métricas de performance utilizadas foram acurácia (83%), precisão (70%), recall (55%) e AUC (88%);
2. O modelo possui uma boa precisão ao classificar pessoas que possuem renda anual > $50,000 USD (70%). Contudo, a maior parte das classificações erradas feitas por este modelo são Falsos Negativos (indivíduos com renda anual > $50,000 USD, que são classificados como possuindo renda anual < $50,000 USD;
3. Aviso: a performance do modelo varia consideravelmente entre subgrupos de atributos sensíveis (e.g., gênero, raça, estado civil);
4. Dados de treinamento e testagem foram adquiridos diretamente do conjunto de dados fornecidos pela UCI Machine Learning Repository (i.e., Adult Census Income Data Set);
5. Este conjunto de dados foi escolhido por sua disponibilidade pública;
6. Amostras com valores ausentes (i.e., “?” ou “NaN”) foram excluídas do conjunto de dados.

Considerações Éticas

1. Dada a distribuição enviesada dos dados de treinamento, o modelo pode se comportar de forma ineficiente quando lidando com amostras pouco vistas. Sua performance varia consideravelmente entre subgrupos, não alcançado padrões mínimos de poder preditivo para certos subgrupos (e.g., Casado-cônjuge-militar);
2. Recomenda-se que para aplicações reais, o conjunto de dados seja ampliado, de forma que aja uma melhor distribuição de amostras por subgrupos de características;
3. De acordo com os resultados de performance e matrizes de confusão entre subgrupos, atributos sensíveis podem interferir na predição deste modelo.

Detalhes e Recomendações

1. O modelo treinado resulta em uma performance que varia entre subgrupos pertencentes a características/atributos sensíveis. Se utilizado para aplicações que podem causar impacto na vida das pessoas (e.g., determinando que deve pagar impostos mais elevados), o modelo pode vir a prejudicar populações sub representadas no Adult Census Income Data Set;
2. Os dados utilizados para este exemplo não refletem o contexto social e histórico de um lugar como, por exemplo, Brasil. Eles refletem o contexto social e histórico

Norte-Americano. Assim, não se recomenda utilizá-lo para desenvolvimento de aplicações fora deste domínio específico.

Análise Quantitativa

Performance por Gênero

	Acurácia	Precisão	Recall	AUC
Masculino	0.7888	0.7492	0.4917	0.8592
Feminino	0.9145	0.7363	0.3867	0.9120

Performance por Raça

	Acurácia	Precisão	Recall	AUC
Caucasiano	0.8227	0.7527	0.4882	0.8812
Negro	0.8896	0.7068	0.2568	0.9102
Asiático-Americano	0.7966	0.6774	0.5081	0.8592
Esquimó	0.8951	0.6429	0.2647	0.7831
Outro	0.9134	0.5385	0.3333	0.9209

Performance por Estado Civil

	Acurácia	Precisão	Recall	AUC
Casado (cônjuge civil)	0.7120	0.7475	0.5541	0.7900
Divorciado	0.8949	0.7143	0.0332	0.7959
Casado (cônjuge ausente)	0.9189	0.6667	0.0645	0.8214
Nunca Casado	0.9524	1.0000	0.0149	0.8859
Separado	0.9329	0.8000	0.0606	0.8442
Casado (cônjuge militar)	0.5238	0.0000	0.0000	0.6955
Viúvo	0.9033	0.5000	0.0125	0.7569

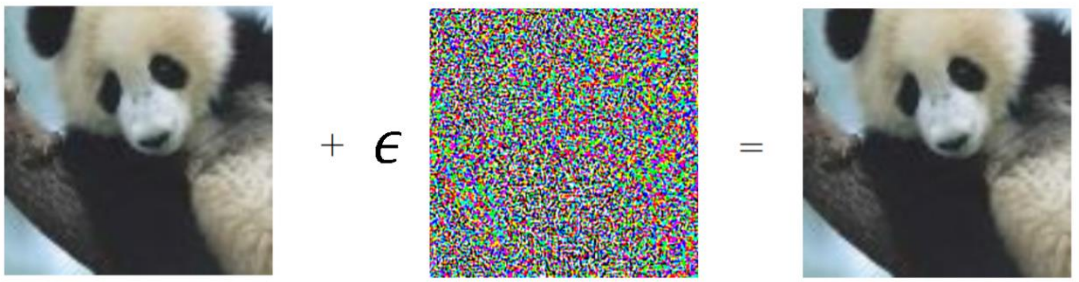
	Acurácia	Precisão	Recall	AUC
Performance do Preditor de Renda Anual	0.8325	0.7074	0.5577	0.8832

Esperamos que os exemplos (assim como as ferramentas) apresentados neste trabalho possam auxiliar desenvolvedores a conceber e melhorar suas próprias análises de segurança, instituindo assim a Ética e Segurança da IA como uma parte integral do processo de desenvolvimento de sistemas inteligentes. Na próxima seção, apresentaremos uma última metodologia a ser incorporada em uma análise de segurança: *ataques adversariais.*

Ataques Adversariais

Modelos criados por aprendizagem de máquina são sistemas curiosos. Por mais que tais sistemas sejam capazes de realizar tarefas extremamente complexas, para as quais não saberíamos como "escrever uma solução", seu funcionamento e a forma como tais sistemas "percebem" o ambiente (i.e., suas entradas), permitem que esses sejam enganados por aquilo que chamamos de "ataques adversariais".

Ataques, ou exemplos, adversariais são entradas/inputs para modelos de aprendizagem de máquina criados com o expresso intuito de fazer com que um modelo cometa um erro (e.g., uma classificação errada) (Szegedy et al., 2013). Esses ataques usam do fato de que modelos de aprendizagem de máquina são (basicamente) conjuntos de funções de ativação e parâmetros otimizados por gradiente descendente. Se tivermos acesso direto (ou indireto) aos valores dos parâmetros de um modelo (ou o próprio gradiente do modelo), podemos utilizar tal informação para corromper sinais de entrada, adicionando perturbações quase que imperceptíveis, para fazer o modelo produzir a saída que nós queremos.



"Panda"
57.7% de confiança

Perturbação

"Gibão"
99.3% de confiança

Um exemplo adversarial, criado ao se adicionar uma pequena perturbação (ϵ) a imagem de um "Panda", para fazer com que uma CNN o classifique como um "Gibão" (Goodfellow et al., 2014, p. 3).

No exemplo acima, Goodfellow et al. (2014) utilizaram conhecimento do gradiente do modelo para criar um exemplo que, para nós, é claramente um panda, mas para o modelo, é um Gibão com 99.3% de confiança. Em outras palavras, os autores avaliaram como a classe "Panda" está próxima da classe "Gibão" dentro do espaço de representações do modelo, e "empurraram" (i.e., perturbaram) tal imagem para fazer com que as representações/parâmetros associados com a classificação da classe "Gibão" fossem fortemente (99.3%) ativados, causando uma classificação errada.

Com exemplos adversariais, atacantes podem explorar potenciais falhas de modelos treinados por aprendizagem de máquina, algo que torna tais entidades dignas de atenção e monitoramento. Por exemplo, Papernot et al. (2016a) demonstraram como imagens de placas de trânsito (e.g., PARE) podem ser alteradas para produzir classificações erradas (e.g., SIGA), algo que poderia vir a causar acidentes de trânsitos envolvendo carros autônomos guiados por visão computacional. Ahmad et al. (2021) sugerem que sistemas de reconhecimento facial utilizados para delimitar acesso a zonas restritas, poderiam ser enganados (e.g., o atacante pode descobrir uma espécie de maquiagem/pintura facial que produz um sinal de reconhecimento com alta confiança) a liberar a entrada de pessoas não autorizadas.

Utilizando como exemplo o modelo para aprovação de cartões de crédito (Exemplo 1), uma forma simples de (i) entender o funcionamento do modelo e (ii) explorá-lo, é falsificando sinais (i.e., criando exemplos adversariais). As entradas do modelo utilizado no Exemplo 1 são apenas tensores de Rank-1 (i.e., vetores com 15 valores de características). Assim podemos criar dois tensores de Rank-1 (com a dimensão apropriada) para testar como o modelo responde. Utilizemos dois exemplos extremos, i.e., onde todos os valores são 0 ou 1:

- `Caso_extremo_1 = np.array([[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]);`
- `Caso_extremo_2 = np.array([[1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]]).`

Os sinais/amostras "normais" possuem valores muito mais variados (e.g., `amostra[10] = ([[0.0, 0.84393064, 0.02982143, 0.5, 0.0, 0.0, 0.88888889, 0.01754386, 0.0, 0.0, 0.0, 1.0, 0.0, 0.49112426, 0.00228963]])`). Como o modelo criado responde a este tipo de entrada?

	Aprovado	**Reprovado**
Caso_extremo_1	0.20780003	0.79219997
Caso_extremo_2	0.99278847	0.00721153

Já sabemos agora que amostras contendo vários zeros irão gerar reprovações (com 79% de confiança), e que amostras contendo vários uns irão gerar aprovações (com 99% de confiança). Tudo que precisaríamos fazer agora é modificar (sutilmente) os valores de entrada para criarmos inúmeros exemplos de amostras que serão classificadas da forma como desejarmos.

Na batalha entre atacantes e defensores, os defensores encontram-se em desvantagem. Exemplos adversariais não são necessariamente soluções inválidas, mas sim "soluções inesperadas" para um problema de otimização complexo. Nós utilizamos aprendizagem de máquina para encontrar soluções para problemas que nós mesmos não sabemos como solucionar de forma direta. Como muitos dos processos que guiam a otimização de problemas não-lineares e não-convexos ainda não são totalmente compreendidos (Como a inicialização aleatória dos parâmetros de uma rede neural pode influenciar a sua performance final?) (Frankle & Carbin, 2019), nós não possuímos nenhum teorema ou

garantia formal que nos permita detectar/excluir/proteger um modelo contra exemplos adversariais.

Assim, defensores não possuem ferramentas para proteger um modelo contra todos os tipos de ataques possíveis, pois nós não sabemos como encontrá-los de forma sistemática e completa. Enquanto isso, os atacantes apenas precisam encontrar "uma falha". Uma perturbação que os aproxime do resultado desejado. E como isso, dobrar o modelo a sua vontade. Projetar defesas contra ataques adversariais segue sendo um problema em aberto na segurança da IA.

> O estudo de exemplos adversariais é empolgante porque muitos dos problemas mais importantes permanecem abertos, tanto em termos teóricos como em termos de aplicações. Do lado teórico, ninguém ainda sabe se a defesa contra exemplos adversariais é um esforço teoricamente sem esperança (como tentar encontrar um algoritmo universal de aprendizagem de máquina) ou se uma estratégia ótima daria ao defensor alguma vantagem (como na criptografia e na privacidade diferencial). No lado aplicado, ainda ninguém projetou um algoritmo de defesa verdadeiramente poderoso que pudesse resistir a uma grande variedade de algoritmos de ataque de exemplos adversariais (Goodfellow & Papernot, 2017).

Existem uma série de benchmarks para avaliação de robustez de modelos, por onde podemos realizar testes de estresse, e encontrar situações onde nossos modelos falharam (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021b; Koh et al., 2021). Assim, algo que um engenheiro de segurança em aprendizagem de máquina pode fazer é se tornar o primeiro "atacante" de seu próprio modelo. Ou seja, administrar ataques adversariais deve ser uma das etapas essenciais de desenvolvimento e monitoramento de um modelo, antes e após sua implantação.

Grande parte da pesquisa atual em ataques adversariais se concentra no problema de "l_p adversarial robustness", i.e., situações onde atacantes

buscam induzir um modelo ao erro, mas limitando as perturbações introduzidas na amostra dentro de uma pequena restrição ("pequenas perturbações") (Carlini & Wagner, 2017). Ataques podem ser construídos com base em informações internas do modelo (e.g., seu gradiente/valores de parâmetros, como foi feito no exemplo do "Panda/Gibão"), ou apenas através da observação da relação entre entradas/saídas do modelo (e.g., como foi demonstrado no exemplo de aprovação de cartões de crédito) (Tramèr et al., 2018).

Existem diversas estratégias para se desenvolver exemplos adversariais, como busca por força bruta (i.e., geração massiva de exemplos para se encontrar amostras adversariais), geração de dados artificiais/*data augmentation* (Engstrom et al., 2020; Zhu et al., 2021; Rebuffi et al., 2021), e técnicas de aprendizagem que beneficiam a detecção de amostras fora da distribuição de treinamento e amostras anômalas/outliers difíceis de se classificar (e.g., *self-supervised learning*) (Hendrycks et al., 2019).

Aos interessados em aprender mais sobre técnicas de construção de exemplos adversariais, CleverHans[45] é uma biblioteca de software que fornece implementações de referência padronizadas de modo a auxiliar desenvolvedores a criar modelos mais robustos a amostras adversariais. Usando CleverHans, desenvolvedores podem criar seus próprios conjuntos de dados adversariais, de forma padronizada, e treinar seus modelos para tratar tais amostras de forma robusta. Desenvolvedores podem até mesmo criar seus próprios benchmarks de avaliação/treinamento contra exemplos adversariais (Papernot et al., 2016b).

Ian Goodfellow e Nicolas Papernot (criadores da biblioteca CleverHans) mantém um blog [46] sobre segurança e privacidade na aprendizagem de máquina. Nele, é possível encontrar exemplos comentados, junto com

[45] Disponível em: https://Github.com/cleverhans-lab/cleverhans.

[46] Disponível em: http://www.cleverhans.io/.

scripts de código-aberto, ensinando desenvolvedores sobre como realizar análises de segurança.

Considerações Finais

É importante ressaltar que até o presente momento, existem poucas evidências de que o uso de qualquer uma das ferramentas/métodos mencionados neste guia sejam eficientes para otimizar o design ético de sistemas algorítmicos. Como tal, ainda é necessário que estudos que visem a implementação destas técnicas demonstrem os resultados de suas metodologias, seja auxiliando grupos sociais desfavorecidos, ou evitando possíveis efeitos colaterais de sistemas de IA mal concebidos.

O objetivo principal deste guia é munir desenvolvedores de sistemas de IA com ferramentas e métodos para serem aplicados durante o ciclo de vida desses tipos de sistemas. Será apenas através da experimentação que poderemos saber quais ferramentas funcionam, quais funcionam melhor, e quais devem ser melhoradas.

Esperamos ter auxiliado a todos os interessados em diminuir a lacuna entre teoria e prática do desenvolvimento ético e seguro da IA há ampliar seus conhecimentos.

The AI Robotics Ethics Society®

Referências

Agüera y Arcas, B., Todorov, A., & Mitchell, M. (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? *Medium*. https://link.medium.com/GO7FJgFgM1.

Ahmad, Z., Khan, A. S., Shiang, C. W., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32(1), e4150. doi: 10.1002/ett.4150.

AI Robotics Ethics Society (AIRES) at PUCRS. (2021). An Open Letter to the Global South: Bring the "rest" in. AI Robotics Ethics Society.

AlgorithmWatch. (2020). AI Ethics Guidelines Global Inventory. Algorithm Watch. https://inventory.algorithmwatch.org/.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. ArXiv. https://arxiv.org/abs/1606.06565.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. doi:10.1177/1461444816676645.

Badia, A., Bilal, P., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, D., & Blundell, C. (2020). Agent57: Outperforming the Atari Human Benchmark. DeepMind. https://arxiv.org/pdf/2003.13350.pdf.

Balch, O. (2020). AI and me: friendship chatbots are on the rise, but is there a gendered design flaw? *The Guardian*. https://www.theguardian.com/careers/2020/may/07/ai-and-me-friendship-chatbots-are-on-the-rise-but-is-there-a-gendered-design-flaw.

Baum, S. (2017). A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute, Working Paper, 1-17. http://dx.doi.org/10.2139/ssrn.3070741.

Bender, E. M., & Friedman, B. (2018). Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics,* 6, 587-604. doi: 10.1162/tacl_a_00041.

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer International Publishing. doi: 10.1007/978-3-319-60648-4.

Bonilla-Silva, E. (2013). *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States* (4th edition). Rowman & Littlefield Publishers.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. ArXiv. https://arxiv.org/pdf/2005.14165.pdf.

Calo, R. (2017). Artifcial intelligence policy: a primer and roadmap. *SSRN Journal*, 399–435. doi:10.2139/ssrn.3015350.

Calvo R. A., Peters D., Vold K., & Ryan R. M. (2020) *Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry. In Ethics of Digital Well-Being, Philosophical Studies Series*, vol 140, Burr C., & Floridi L. (eds.). Springer, Cham. doi: 10.1007/978-3-030-50585-1_2.

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In the *2017 IEEE Symposium on Security and Privacy*. https://arxiv.org/abs/1608.04644.

Carrillo, R. M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, 44(6), 101937. doi: 10.1016/j.telpol.2020.101937.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H., Kaplan, J., Edwards, H., Burda, Y., Joseph, N. et al. (2021). Evaluating Large Language Models Trained on Code. OpenAI. https://arxiv.org/abs/2107.03374.

Chouldechova, A. (2016). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2), 153-163. doi:10.1089/big.2016.0047.

Churchland, P. S., & Sejnowski, T. (1992). *The computational brain*. USA, Cambridge: MIT Press.

Collins, E. (2018). Punishing Risk. *Geo. L. J*, 57. https://ssrn.com/abstract=3171053

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. In *the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. doi:10.1145/3097983.3098095.

Corrêa, N. K., & De Oliveira, N. (2021). Good AI for the Present of Humanity Democratizing AI Governance. *AI Ethics Journal*, 2(2)-2. doi: 10.47289/AIEJ20210716-2.

Critch, A., & Krueger, D. (2020). AI Research Considerations for Human Existential Safety (ARCHES). ArXiv. https://arxiv.org/abs/2006.04948.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zeme, R. (2012). Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference,* 214–226. doi: 10.1145/2090236.2090255.

Ekstrand, M.D., Joshaghani, R., & Mehrpouyan, H. (2018). Privacy for all: Ensuring fair and equitable privacy protections. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 1–13.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Steinhardt, J., & Madry, A. (2020). Identifying Statistical Bias in Dataset Replication. In 2020 *International Conference on Machine Learning.* https://arxiv.org/abs/2005.09619.

Everitt, T., Kumar, R., Krakovna, V., Legg, S. (2019). Modeling AGI Safety Frameworks with Causal Influence Diagrams. DeepMind. https://arxiv.org/abs/1906.08663.

Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. In the *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63. doi: 10.1145/3375627.3375828.

Fitzgerald, M., Boddy, A., & Baum, S. B. (2020). 2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Technical Report 20-1.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence. Mapping Consensus in Ethical and Rights-Based Approaches to Principles for Ai. In Berkman Klein Center Research Publication 2020, p. 1–39.

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, *Physical and Engineering Sciences*, 374(2083), 20160360. doi: 10.1098/rsta.2016.0360.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines*, 28, 689–707. doi:10.1007/s11023-018-9482-5.

Frankle, J., & Carbin, M. (2019). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In the *International Conference on Learning Representations (2019).* https://openreview.net/pdf?id=rJl-b3RcF7.

Fryer, R., Loury, G., & Yuret, T. (2008). An Economic Analysis of Color-Blind Affirmative Action. *Journal of Law, Economics, and Organization*, 24(2), 319–355.

Gajane, P., & Pechenizkiy, M. (2018). On Formalizing Fairness in Prediction with Machine Learning. Department of Computer Science, Montanuniversitat Leoben, Austria, and the Department of Computer Science, TU Eindhoven, the Netherlands. ArXiv. https://arxiv.org/abs/1710.03184.

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness Testing: Testing Software for Discrimination. In *the Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 498–510. doi:10.1145/3106237.3106277.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for Datasets. ArXiv. https://arxiv.org/abs/1803.09010.

Goldsmith, J., & Burton, E. (2017). Why teaching ethics to AI practitioners is important. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4863–4840. https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14271/13992.

Goodfellow, I. J., Shlens, J., Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. In the *2015 International Conference on Learning Representations*. https://arxiv.org/abs/1412.6572.

Goodfellow, I., & Papernot, N. (2017). Is attacking machine learning easier than defending it? Cleverhans-blog. www.cleverhans.io/security/privacy/ml/2017/02/15/why-attacking-machine-learning-is-easier-than-defending-it.html.

Green, B. (2019). "Good" isn't good enough. In *NeurIPS workshop on AI for social good*. https://www.benzevgreen.com/wp-content/uploads/2019/11/19-ai4sg.pdf.

Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., & Williams, B. (2016). Embedding Ethical Principles in Collective Decision Support Systems. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16). https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12457.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv. CSUR*, 51(5), 93:1–93:42. doi: 10.1145/3236009.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30, 99–120. doi:10.1007/s11023-020-09526-7.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *the Proceedings of the 2016 Advances in neural information processing systems*, 29, 3315–3323.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In the *2019 Proceedings of the International Conference on Learning Representations*. https://arxiv.org/abs/1903.12261.

Hendrycks, D., Mazeika, M., Kadavath, S., & Song, D. (2019). Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In the 2019 *Conference on Neural Information Processing Syste*ms. https://arxiv.org/abs/1906.12340.

Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J. (2021a). Unsolved Problems in ML Safety. ArXiv. https://arxiv.org/abs/2109.13916#.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., & Gilmer, J. (2021b). The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In the *2021 International Conference on Computer Vision.* https://arxiv.org/abs/2006.16241.

Hirose, I. (2014). *Egalitarianism* (1st edition). UK, London: Routledge.

Hofstede, G. H., Hofstede, G. J., & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind* (3rd edition). New York, NY: McGraw-Hill.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. ArXiv. https://arxiv.org/abs/1805.03677#.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. Machine Intelligence Research Institute. https://arxiv.org/abs/1906.01820.

Hutter, M. (2005). Universal artificial intelligence: Sequential decisions based on algorithmic probability. *Springer-Verlag Berlin Heidelberg*. doi:10.1007/b138233.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat Mach Intell*, 1, 389–399. doi:10.1038/s42256-019-0088-2.

Jurić, M., Šandić, A., & Brcic, M. (2020). AI safety: state of the field through quantitative lens. *43rd International Convention on Information and Communication Technology, Electronics and Microelectronics* (MIPRO). https://arxiv.org/ftp/arxiv/papers/2002/2002.05671.pdf.

Kärkkäinen, K., & Joo, J. (2019). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. University of California, Los Angeles. ArXiv. https://arxiv.org/abs/1908.04913.

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., Irving, G. (2021). Alignment of Language Agents. DeepMind. https://arxiv.org/abs/2103.14659.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding Discrimination Through Causal Reasoning. In *the Proceedings of the 31st International Conference on Neural Information Processing Systems*, 656–666. doi:10.5555/3294771.3294834.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. Cornell University and Harvard University. ArXiv. https://arxiv.org/abs/1609.05807.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., & Liang, P. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts. In the *2021 International Conference on Machine Learning.* https://arxiv.org/abs/2012.07421.

Krafft, T. D., & Zweig, K. A. (2019). Transparency and traceability of algorithm-based decision-making processes | A regulatory proposal. Verbraucherzentrale Bundesverband (Federal Association of Consumer Organizations). https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf.

Krafft. T. B., Hauer, M., Fetic, L., Kaminski, A., Puntschuh, M., Otto, P., Hubig. C., Fleischer, T., Grünke, P., Hillerbrand, R., Husted. C., & Hallensleben, S. (2020). From Principles to Practice - An interdisciplinary framework to operationalise AI ethics. AI Ethics Impact Group (VDE Association for Electrical, Electronic & Information Technologies/Bertelsmann Stiftung). https://www.ai-ethics-impact.org/en.

Krishnan, M. (2019). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy and Technology*, 33(1). doi: 10.1007/s13347-019-00372-9.

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., Legg, S. (2017). AI Safety Gridworlds. DeepMind. https://arxiv.org/abs/1711.09883.

Lohr, S. (2018). Facial Recognition Is Accurate, if You're a White Guy. *The New York Times*. https://www.nytimes.com/2018/02/09/technology/facialrecognition-race-artificial-intelligence.html.

Luengo-Oroz, M. (2019). Solidarity should be a core ethical principle of AI. *Nat Mach Intell*, 1(494). doi:10.1038/s42256-019-0115-3.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN As an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 502–510.

Maxmen, A. (2018). Self-driving car dilemmas reveal that moral choices are not universal. *Nature*, 562 (7728), 469–470. doi:10.1038/d41586-018-07135-0.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35. doi:10.1145/3457607.

Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., Ortega, P. A. (2020). Meta-trained agents implement Bayes-optimal agents. DeepMind. https://arxiv.org/abs/2010.11223.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P. Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. In the *Proceedings of the Conference on Fairness, Accountability, and Transparency* (January, 2019), 220–229. doi:10.1145/3287560.3287596.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency,* 279–288. doi:10.1145/3287560.3287574.

Mnhi, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. DeepMind. https://arxiv.org/abs/1312.5602.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168. doi: 10.1007/s11948-019-00165-5.

Morley, J., Elhalal, A., Garcia, F., Kinsey, l., Mökander, J., & Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI Ethics. *Minds and Machines*, 31,239–256. doi:10.1007/s11023-021-09563-w.

Newell, A. (1990). *Unified theories of cognition*. USA, Cambridge: Harvard University Press.

Nunes, P. (2019). EXCLUSIVO: levantamento revela que 90,5% dos presos por monitoramento facial no Brasil são negros. *The Intercept Brasil*. https://theintercept.com/2019/11/21/presos-monitoramento-facial-brasil-negros/.

Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A., Matyasko, A., Behzadan, V., Hambardzumyan, K., Zhang, Z., Juang, Y., Li, Z., Sheatsley, R., Garg, A., Uesato, J., Gierke, W., Dong, Y., Berthelot, D., Hendricks, P., Rauber, J., Long, R., & McDaniel, P. (2016a). Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. ArXiv. https://arxiv.org/abs/1610.00768.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, B. Z., & Swami, A. (2016b). Practical Black-Box Attacks against Machine Learning. In the *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, UAE.* https://arxiv.org/abs/1602.02697.

Pearl, J. (1995). *Causation, Action, and Counterfactuals*. In *Computational Learning and Probabilistic Reasoning*, A. Gammerman (ed.), USA, New York: John Wiley and Sons, 235–255.

Rahimi, A [Preserve Knowledge]. (2018, March 7). NIPS 2017 Test of Time Award "Machine learning has become alchemy." | Ali Rahimi, Google [Video]. Youtube. https://www.youtube.com/watch?v=x7psGHgatGM.

Rawls, J. (1999). A Theory of Justice. UK, Oxford: Oxford University Press.

Rebuffi, S., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Fixing Data Augmentation to Improve Adversarial Robustness. In the *2021 Conference on Neural Information Processing Syste*ms. https://arxiv.org/abs/2103.01946.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 1-5. doi:10.1177/2053951720942541.

Russell, S., Dewey, D., & Tegmark, M. (2015). An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence. Open Letter. Signed by 8,600 people. https://futureoflife.org/data/documents/research_priorities.pdf.

Ruster, L. (2021). Dignity & Artificial Intelligence: Exploring the role of dignity in government AI ethics instruments. Centre for Public Impact. https://www.centreforpublicimpact.org/partnering-for-learning/cultivating-a-dignity-ecosystem-in-government-ai-ethics-instruments.

Saravanakumar, K. K. (2021). The Impossibility Theorem of Machine Fairness -- A Causal Perspective. Columbia University. ArXiv. https://arxiv.org/abs/2007.06024.

Sen, A. (1990). Justice: Means versus Freedoms. *Philosophy and Public Affairs*, 19.

Silver, D., Huang, A., Maddison, C., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484–489. doi:10.1038/nature16961.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, 299(103535). doi:10.1016/j.artint.2021.103535.

Soares, N. (2016). Value Learning Problem. In *Ethics for Artificial Intelligence Workshop, 25th International Joint Conference on Artificial Intelligence* (IJCAI-2016), USA, New York 9–15. https://intelligence.org/files/ValueLearningProblem.pdf.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *Artificial Intelligence and Ethics*, T. Walsh (ed.), AAAI Technical Report WS-15-02. Palo Alto, CA: AAAI Press.

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle. *MIT Case Studies in Social and Ethical Responsibilities of Computing*. doi:10.21428/2c646de5.c16a07bb.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2013). Intriguing properties of neural networks. In the *2014 International Conference on Learning Representations*. Intriguing properties of neural networks. https://arxiv.org/abs/1312.6199.

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & Mcdaniel, P. (2018). Ensemble Adversarial Training: Attacks and Defenses. In the *2018 International Conference on Machine Learning.* https://arxiv.org/abs/1705.07204.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *the Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7.

Wang, Y., & Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. doi:10.1037/pspa0000098.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, 3(160018). doi:10.1038/sdata.2016.18.

Wolf, M., Miller, K., & Grodzinsky, F. (2017). Why we should have seen that coming: comments on microsoft's tay experiment, and wider implications. *ACM SIGCAS Computers and Society*, 47(3), 54–64.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., & Gao, I. (2021). Mastering Atari Games with Limited Data. In the *35th Conference on Neural Information Processing Systems (NeurIPS 2021).* https://arxiv.org/abs/2111.00210.

Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 30(9), 2805–2824. doi:10.1109/TNNLS.2018.2886017.

Zhu, Y., Ma, J., Sun, J., Chen, Z., Jiang, R., & Li, Z. (2021). Towards Understanding the Generative Capability of Adversarially Robust Classifiers. In the *2021 International Conference on Computer Vision.* https://arxiv.org/abs/2108.09093.