Capítulo 8

“A Nova Eletricidade”: Aplicações, Riscos e Tendências da IA Moderna

Ana L. C. Bazzan, Anderson R. Tavares, André G. Pereira, Cláudio R. Jung, Jacob Scharcanski, Joel Carbonera, Luis C. Lamb, Mariana Recamonde-Mendoza, Thiago L. T. da Silveira, Viviane Moreira[1]

Programa de Pós-Graduação em Computação (PPGC) - UFRGS

Resumo

A provocativa comparação entre IA e eletricidade, feita pelo cientista da computação e empreendedor Andrew Ng, resume a profunda transformação que os recentes avanços em Inteligência Artificial (IA) têm desencadeado no mundo. Este capítulo apresenta uma visão geral pela paisagem em constante evolução da IA. Sem pretensões de exaurir o assunto, exploramos as aplicações que estão redefinindo setores da economia, impactando a sociedade e a humanidade. Analisamos os riscos que acompanham o rápido progresso tecnológico e as tendências futuras da IA, área que trilha o caminho para se tornar uma tecnologia de propósito geral, assim como a eletricidade, que revolucionou a sociedade dos séculos XIX e XX.

Abstract

The provocative comparison between AI and electricity, made by computer scientist and entrepreneur Andrew Ng, summarizes the deep transformation that recent advances in Artificial Intelligence (AI) have triggered in the world. This chapter provides an overview of the ever-evolving landscape of AI. Without intending to exhaust the subject, we explore the applications that are redefining sectors of the economy, impacting society and humanity. We analyze the risks that accompany rapid technological progress and future trends in AI, an area that is on the path to becoming a general-purpose technology, just like electricity, which revolutionized society in the 19th and 20th centuries

Vídeo com a apresentação do capítulo: `https://youtu.be/_1rtWWHFjdw`

[1]Lista de autores em ordem alfabética.

Parte I: Introdução e Fundamentos

8.1. Introdução

Comparar a Inteligência Artificial (IA) e a eletricidade foi a maneira que o cientista da computação e empreendedor Andrew Ng usou para sintetizar o potencial transformador e também os perigos dessa tecnologia [Lynch 2017]. Assim como a eletricidade moldou a história do século XIX, a IA está esculpindo o cenário do século XXI de maneiras que desafiam as fronteiras do conhecimento e da imaginação, com um farol de possibilidades intrigantes e, ao mesmo tempo, profundas preocupações sobre os rumos que a tecnologia pode tomar, mesmo quando usada para fins não-maliciosos.

Indo além da analogia IA-eletricidade de Andrew Ng, o status da IA como uma força transformadora foi reconhecido pela comunidade científica da Ciência da Computação. O prestigiado Prêmio Turing, também chamado de "Nobel da Computação", foi concedido em 2018 aos cientistas da computação Yoshua Bengio, Geoffrey Hinton e Yann LeCun por suas contribuições pioneiras para o desenvolvimento da aprendizagem profunda, o componente da IA no cerne da disrupção provocada na sociedade. Esse reconhecimento não apenas consagrou a significância da IA na era contemporânea, mas também destacou o papel crucial desses visionários em pavimentar o caminho para avanços que reverberam em todas as facetas da sociedade.

Este capítulo introduz conceitos básicos de IA (Seção 8.2) e, na Parte II, apresenta uma visão geral de suas implicações em diversas áreas: Visão Computacional (Seção 8.3), Processamento de Linguagem Natural (Seção 8.4), Saúde (Seção 8.5), Indústria (Seção 8.6), Finanças (Seção 8.7) e Mobilidade Urbana (Seção 8.8). Usando aspectos da provocativa analogia de Ng, em cada área serão apresentadas algumas aplicações, riscos e tendências. A Parte III apresentam um panorama geral do trabalho, revisitando riscos em comum nas diferentes áreas; discute a IA neuro-simbólica como uma abordagem promissora pra esses riscos; e conclui com um chamado à reflexão sobre os rumos da tecnologia e da humanidade. O capítulo não tem pretensões de esgotar o assunto, nem na listagem das áreas impactadas pela IA, pois praticamente todos os aspectos da vida em sociedade serão afetados, nem nas aplicações específicas de cada área. O conteúdo aqui apresentado é um convite para jornadas mais abrangentes e profundas do leitor, que poderá expandir seus horizontes nas referências apresentadas.

8.2. Fundamentos

O psicólogo e economista Daniel Kahneman, ganhador de um Prêmio Nobel de Economia, propôs que o raciocínio humano é dividido em dois sistemas [Kahneman 2011]. No Sistema 1, a mente trabalha de maneira instintiva, rápida, por reflexos, sujeita a erros e de maneira difícil de descrever (sem transparência). Este sistema é mais ativo em decisões rotineiras e tarefas mundanas, como os movimentos corretos a serem feitos enquanto se dirige um carro. No Sistema 2, a mente trabalha de maneira deliberada, lenta, confiável e transparente. Este sistema é mais ativo em decisões estratégicas e tarefas intelectuais, como a escolha de rotas enquanto se dirige um carro.

O modelo da mente humana dividida em Sistemas 1 e 2 é também útil para mapear

abordagens de inteligência artificial [Geffner 2018]. Abordagens baseadas em aprendizado, mais modernas em IA, estão mais próximas ao Sistema 1: modelos treinados dão respostas rápidas, sujeitas a erros e difíceis de rastrear (baixa transparência). Abordagens de IA simbólica, uma tradição dominante historicamente em IA até o início dos anos 2000, estão associadas ao Sistema 2: trata-se de métodos e algoritmos que enumeram explicitamente possíveis soluções para um problema e as investigam de maneira sistemática, sendo lentos, porém fáceis de rastrear (transparentes), pois é possível verificar o estado de um algoritmo e entender as decisões feitas por ele. Cabe ressaltar, no entanto, que Kahneman declarou que o Sistema 1 e Sistema 2 atuam de forma integrada, em debate sobre tendências em IA durante a conferência AAAI-2020 [2]. Assim, embora exista a distinção entre Sistemas 1 e 2, os mesmos podem ser vistos de forma harmônica, o que nos remete à IA neuro-simbólica [d'Avila Garcez and Lamb 2023], que analisaremos na Seção 8.10.

O restante desta seção apresenta fundamentos de IA simbólica (Seção 8.2.1) e IA baseada em aprendizado (Seção 8.2.2). Essa organização segue a ordem baseada na linha de tempo da IA, onde os métodos simbólicos (análogos ao Sistema 2 da mente humana) eram tradicionalmente predominantes, enquanto os métodos baseados em aprendizado (análogos ao Sistema 1 da mente humana) ganharam proeminência em tempos modernos, sendo responsáveis pela notoriedade atual da IA. Ao leitor interessado em se aprofundar nos conceitos apresentados aqui, o abrangente livro de [Russell and Norvig 2020] é uma excelente referência para os conceitos de IA em geral.

8.2.1. IA simbólica

Representação de conhecimento e raciocínio é uma área da IA preocupada com como o conhecimento pode ser representado de forma simbólica e manipulado de maneira automatizada por programas que representam processos de raciocínio realizados sobre as representações simbólicas [Brachman and Levesque 2004]. Esta abordagem da IA busca estudar e produzir comportamento inteligente sem considerar necessariamente a estrutura biológica subjacente ao processo de raciocínio, mas focando no conhecimento que os agentes possuem. Assim, parte-se do pressuposto que o que permite aos agentes (como humanos) se comportarem de maneira inteligente é que eles sabem muitas coisas sobre o seu ambiente e são capazes de aplicar esse conhecimento conforme necessário para se adaptar às situações que se apresentam para alcançar seus objetivos. Portanto, nesta abordagem de IA, focamos no conhecimento e na representação deste conhecimento. As questões-chave nesta área são o que qualquer agente (humano, animal, artificial, etc) precisaria saber para se comportar de maneira inteligente e que tipos de mecanismos computacionais poderiam permitir que seu conhecimento fosse manipulado para realizar inferências que possibilitem que o agente atinja seus objetivos [Fagin et al. 1995].

Na área de representação de conhecimento e raciocínio, o conhecimento é geralmente visto como uma coleção de proposições, que são formulações abstratas geralmente representadas por sentenças declarativas sobre o mundo (ou alguma parte ou aspecto dele) que podem ser verdadeiras ou falsas. Nesta área da IA, em geral, proposições são representadas por símbolos (como sequências de caracteres com sintaxe bem definida), que, ao

[2] AAAI-2020 Fireside Chat with Daniel Kahneman - com Francesca Rossi, Yoshua Bengio, Geoff Hinton e Yann LeCun. `https://vimeo.com/390814190` Acesso em 25 de setembro de 2023.

contrário das proposições, são entidades concretas e que permitem a manipulação das proposições por meios computacionais [Kowalski 1979]. Na abordagem simbólica, pesquisadores fazem utilização intensiva de formulações precisas, através das lógicas clássicas (proposicional e de predicados de primeira ordem) e lógicas não-clássicas, e.g. modais, temporais, epistêmicas, espaciais, probabilísticas, entre outras [Broda et al. 2004]. Assim, a representação de conhecimento é o campo de estudo focado em estudar o uso de símbolos formais para representar proposições que constituem o conhecimento de um agente. Neste contexto, raciocínio é a manipulação computacional deste símbolos que representam o conhecimento de um agente, visando produzir novos símbolos, que representam um novo conhecimento. Uma coleção de representações simbólicas de conhecimento constitui o que chamamos de uma *base de conhecimento*. Sistemas cujo comportamento inteligente é produzido pela manipulação das bases de conhecimento através de mecanismos computacionais que descrevem um processo de raciocínio são chamados de *sistemas baseados em conhecimento*. Neste tipo de sistema, adota-se uma abordagem declarativa, em que especificamos o que o agente sabe (onde fatos novos podem ser descobertos via percepção do ambiente) e o agente pode chegar a conclusões (que podem ser ações) através de inferências lógicas realizadas sobre sua base de conhecimento. Na abordagem declarativa de construção de sistemas, não especificamos o fluxo de controle da manipulação de dados, como em abordagens procedimentais.

Embora existam diversas abordagens, tipicamente a construção de sistemas ou agentes *baseados em conhecimento* envolve o uso de linguagens formais baseadas em lógica (como a lógica de primeira ordem e lógica modal, entre outras) para representar conhecimento dependente de tarefa e domínio. Com isso, podemos definir mecanismos de raciocínio independentes de tarefa e domínio, que realizam processos computacionais que representam inferências lógicas bem fundamentadas que garantem propriedades lógicas desejáveis, como a validade. Estes mecanismos de raciocínio, por sua vez, são capazes de derivar novo conhecimento a partir dos fatos armazenados na base de conhecimento. Historicamente, a área de IA simbólica teve grande impulso a partir dos anos 1970 com desenvolvimentos em programação em lógica [Kowalski 1979]. Especificamente, o desenvolvimento da linguagem Prolog (Programming in Logic) permitiu que pesquisadores passassem a expressar conhecimentos de domínios específicos de forma declarativa, sob uma fundamentação lógica e desenvolvessem sistemas computacionais a partir de uma abordagem simbólica. Prolog também teve grande impacto nos anos 1980 e 1990, notadamente durante o projeto liderado pelo Japão, denominado de "Fifth Generation Computer Systems Project". A linguagem Prolog foi aplicada com sucesso na prova automática de teoremas, sistemas especialistas, planejamento, bancos de dados, processamento de linguagem natural, aplicações legais entre outras áreas onde a representação de conhecimento e inferência lógica exigem uma representação computacional adequada [Warren et al. 2023].

Planejamento automatizado é uma subárea da IA simbólica que visa produzir solucionadores com comportamento direcionado a objetivos que sejam gerais e eficientes. Esses solucionadores, também chamados de planejadores, aceitam modelos que visam produzir comportamento direcionado a objetivos, sendo planejamento clássico um dos modelos mais pesquisados. Uma tarefa modelada por planejamento clássico possui um estado inicial, uma condição objetivo e um conjunto de operadores. Uma solução para

uma tarefa de planejamento é uma sequência de operadores que satisfazem a condição objetivo quando aplicados ao estado inicial. Um exemplo de tarefa modelada no planejamento clássico é um cenário em um armazém onde o objetivo é encontrar uma sequência de movimentos que os robôs devem realizar para alcançar suas posições-objetivo.

A motivação do planejamento automatizado é criar um planejador que tenha um bom desempenho em qualquer tarefa sem conhecimento prévio [Hoffmann 2011]. Isso torna o planejamento uma abordagem com bom custo-beneficio para desenvolvimento de soluções. Pode-se construir ou selecionar um planejador, descrever qualquer tarefa no modelo do planejador e então resolvê-la usando o planejador. Se a tarefa mudar, basta mudar o modelo da tarefa, mas não o planejador. Assim, utilizar planeadores para encontrar boas soluções para problemas do mundo real pode ajudar a reduzir tempo e custos.

Os planejadores mais bem-sucedidos baseiam-se em busca heurística. A^* é o algoritmo de busca heurística mais conhecido [Hart et al. 1968]. Ele processa nodos de maneira sistemática, buscando o caminho ótimo (de menor custo) entre o estado inicial e o objetivo, empregando uma função heurística para descartar caminhos não-promissores. Heurísticas bem construídas aumentam a eficiência do algoritmo enquanto mantém sua otimalidade.

Funções heurísticas usam o modelo da tarefa para raciocinar automaticamente sobre a mesma, gerando estimativas do custo para satisfazer a condição objetivo. Em geral, este processo de raciocínio utiliza relaxações ou abstrações para calcular as estimativas em tempo polinomial [Helmert and Domshlak 2009]. Por esta razão, os planejadores são chamados de independentes de domínio. Eles podem calcular estimativas diretamente do modelo sem conhecimento prévio sobre o domínio ou a tarefa.

8.2.2. IA baseada em aprendizado

IA baseada em aprendizado é também conhecida como Aprendizado de Máquina. Em aprendizado de máquina, os sistemas computacionais são *treinados* para realizar uma tarefa, ao invés de serem explicitamente programados para isso, como na IA simbólica. O ponto chave em aprendizado de máquina é que os sistemas computacionais consigam realizar uma tarefa e melhorar seu desempenho nela à medida que adquirem mais dados ou experiência [Mitchell 1997]. Esta seção apresenta uma visão geral da área, dividida nas três subáreas mais comuns: aprendizado supervisionado 8.2.2.1, não-supervisionado 8.2.2.3 e por reforço 8.2.2.4.

8.2.2.1. Aprendizado supervisionado

Aprendizado supervisionado é uma área relacionada a tarefas de predição. Exemplos rotineiros incluem: predizer a probabilidade de chuva a partir dos dados climáticos, qual o valor de um ativo a partir de seus dados históricos, quais objetos estão presentes em uma imagem, qual a próxima palavra a se escrever a partir das palavras já escritas, entre outros. Para se treinar um modelo de aprendizado supervisionado, é necessária a existência de um conjunto de dados rotulados, isto é, composto por instâncias definidas como pares de entrada e saída esperada. Através do seu processo de aprendizado, o algoritmo utilizará estes dados para encontrar padrões que mapeiam cada entrada para sua saída esperada.

Trata-se, portanto, de uma tarefa com *feedback* instrutivo, onde, para cada entrada, há a instrução de qual é a saída ou resposta correta.

Quando a saída esperada no conjunto de dados faz parte de um conjunto finito de possibilidades, a tarefa de aprendizado supervisionado é de *classificação*. Dentre os exemplos do início dessa seção, a predição de quais objetos estão presentes em uma imagem e qual a próxima palavra a se escrever a partir das palavras já escritas são tarefas de classificação. Quando a saída esperada no conjunto de dados é um número, a tarefa de aprendizado supervisionado é de *regressão*. Dentre os exemplos do início dessa seção, a predição da probabilidade de chuva a partir dos dados climáticos e do valor de um ativo a partir de seus dados históricos são tarefas de regressão.

Dentre as categorias de algoritmos de aprendizado supervisionado mais tradicionais, incluem-se (de maneira não-exaustiva):

- Métodos baseados em instâncias, como "k-vizinhos mais próximos", onde não há um modelo treinado, mas um dado recebe a classe de seus vizinhos mais próximos;
- Árvores de decisão, onde o conjunto de dados é sucessivamente particionado de acordo com os valores do atributo escolhido em cada ponto (nó) de decisão;
- Métodos probabilísticos, como Naïve Bayes, onde a probabilidade de uma instância ser de uma dada classe depende das ocorrência de seus atributos em cada classe do conjunto de treino;
- Métodos de combinação de modelos (*ensembles*), onde múltiplos modelos preditivos são combinados para melhor desempenho geral;
- Métodos conexionistas, como as redes neurais, que inspiram-se na capacidade de processamento de sinais do cérebro biológico.

A seguir, apresentamos uma breve descrição de redes neurais, visto que elas tem sido o componente principal dos sistemas mais modernos e disruptivos de IA.

8.2.2.2. Redes neurais e aprendizado profundo

Redes neurais são modelos computacionais inspirados na estrutura e funcionamento do cérebro biológico. De maneira simplificada, os neurônios que compõem o cérebro biológico são elementos capazes de processar sinais elétricos, recebendo sinais de outros neurônios, atenuando-os ou amplificando-os e combinando-os em um sinal de saída a ser processado por outros neurônios. [McCulloch and Pitts 1943] foram pioneiros ao propor uma versão matemática desse elemento: o neurônio artificial recebe "sinais" numéricos como entrada, e combina-os em uma soma ponderada, na qual pesos numéricos fazem o papel de atenuar ou amplificar os números recebidos como entrada. Os pesos são os parâmetros do neurônio e podem ser modificados para melhorar seu desempenho. Uma função de ativação é então aplicada a essa soma ponderada, resultando na saída do neurônio.

Quando os neurônios artificiais são organizados em camadas interconectadas e a função de ativação é não-linear, temos o perceptron multicamada, o modelo mais tradicional de redes neurais. Quando o perceptron multicamada tem pelo menos uma camada intermediária (ou oculta) entre a entrada e a saída, já é possível dizer que trata-se de aprendizado profundo, pois tal rede já é capaz de detectar e combinar características não-lineares nos dados de entrada [Goodfellow et al. 2016], embora alguns pesquisadores somente "reconhecem" como profundas as redes com dezenas de camadas.

O treinamento de uma rede neural é uma forma de otimização baseada em gradiente. Nessa modalidade, uma função de custo, contínua e diferenciável, avalia as diferenças entre predições da rede e saídas esperadas. O gradiente dessa função de custo indica a mudança necessária em cada parâmetro (peso) da rede para que o custo, e consequentemente os erros da rede, diminuam. Em termos práticos, o treino envolve a apresentação dos dados de entrada (números colocados na primeira camada), o cálculo das ativações da camada inicial até a final e a comparação com a saída esperada. Os pesos de todos os neurônios da rede são ajustados na direção determinada pela derivada parcial da função de custo com relação a cada peso. Essa derivada parcial é exatamente a mudança necessária em cada peso para que o custo diminua. Essas derivadas são calculadas da camada final da rede em retropropagação até a camada inicial. Esse procedimento é o clássico algoritmo *backpropagation* de [Rumelhart et al. 1986], cujos princípios são a base de praticamente todos os sistemas de aprendizado profundo.

De uma maneira simplificada, uma rede neural é um conjunto de parâmetros (números) que realiza transformações numéricas em suas entradas. A clássica organização em perceptron multicamadas é uma das formas de se estruturar uma rede. Ela é especialmente eficaz para dados estruturados, com características já extraídas. Porém, para processamento de dados não-estruturados, a extração de características é necessária. Arquiteturas específicas de redes neurais são capazes de fazer essa extração de características, também chamada de aprendizado de representações. A seguir, descrevemos brevemente dois tipos de redes neurais que se tornaram modelos básicos em suas respectivas áreas: redes neurais convolucionais (CNNs do inglês *convolutional neural networks*), extensivamente usadas em visão computacional (ver Seção 8.3) e Transformers, responsáveis por grandes avanços em processamento de linguagem natural (ver Seção 8.4).

Em processamento de imagens, convoluções são operações matemáticas, nas quais filtros (ou *kernels*) percorrem a imagem de entrada, na forma de uma janela deslizante, gerando um mapa de características. Os filtros definem uma característica de interesse a ser detectada e o mapa de característica resultante é uma "imagem", na qual os pixels da imagem original contendo a característica de interesse são destacados e os demais são atenuados. O objetivo é capturar padrões e características específicas, como bordas, texturas e formas, em diferentes partes da imagem. [LeCun et al. 1989] foram pioneiros ao vislumbrar que os filtros convolucionais não precisavam ser pré-definidos, mas poderiam ser aprendidos via *backpropagation* para extraírem características relevantes na tarefa em questão. Redes convolucionais estão no centro de várias aplicações em visão computacional, e o leitor interessado pode encontrar mais detalhes na Seção 8.3.

Textos em linguagem natural são sequências de palavras, caracterizando-se como dados não-estruturados. O aprendizado de representações em texto envolve uma série de

desafios significativos, em grande parte devido à natureza complexa, variável e ambígua da linguagem natural. Notáveis avanços foram feitos com a ideia de mapear uma palavra para um vetor cujas coordenadas no espaço dão uma ideia de significado [Mikolov et al. 2013]. Nessa abordagem, palavras similares ficam em coordenadas próximas e há possibilidade de operações aritméticas entre palavras, como o clássico exemplo onde "rei - homem + mulher = rainha". No entanto, essa abordagem tem a limitação de que palavras com diferentes significados são mapeadas para uma única representação, sendo insuficiente para saber, por exemplo se "banco" se refere à agência bancária ou ao local de se sentar. Transformers [Vaswani et al. 2017] resolvem esse problema com o mecanismo de atenção: a representação de cada palavra não mais é fixa, agora ela depende do contexto (demais palavras anteriores e posteriores). Tal mecanismo de atenção é composto de matrizes numéricas, cujos valores são aprendidos via *backpropagation*. Transformers são o motor dos sistemas de processamento de linguagem natural mais impressionantes da atualizade, como o ChatGPT. A Seção 8.4 apresenta mais informações sobre processamento de linguagem natural.

8.2.2.3. Aprendizado não-supervisionado

Aprendizado não-supervisionado lida com tarefas de descrição, em contraste com aprendizado supervisionado (predição) e por reforço (controle). Tarefas de descrição envolvem a extração de padrões, similaridades e informações ocultas nos dados sem a necessidade de rótulos ou supervisão explícita. De maneira sucinta, os principais tipos de tarefas descritivas são:

- Agrupamento (Clustering): o objetivo é dividir o conjunto de dados em grupos de acordo com medidas de similaridade. Tipos de agrupamento incluem particional, no qual o espaço de estados é dividido em subregiões disjuntas, hierárquico, no qual a divisão particional pode ter múltiplas granularidades, e por densidade, na qual os dados são agrupados de acordo com a densidade (muita aglomeração ou dispersão).
- Associação: o objetivo é identificar conjuntos de itens ou características que ocorrem juntos com alta frequência dentro de um conjunto de dados, revelando relações intrínsecas e estruturas subjacentes. Dentre as aplicações mais recorrentes destas técnicas, estão análise de cestas de compras e recomendação de produtos online.
- Redução de Dimensionalidade: o objetivo é obter uma representação simplificada dos dados de entrada, reduzindo o número de dimensões (equivalente às "colunas" em conjuntos de dados tabulares) onde cada dimensão na nova representação é uma combinação das dimensões da representação original. Técnicas de redução de dimensionalidade são rotineiramente utilizadas para visualização e compressão de dados, sendo parte essencial do *pipeline* de projetos em ciências de dados.

8.2.2.4. Aprendizado por reforço

Aprendizado por reforço (AR) é geralmente associado a tarefas de controle, ou seja, aprender a melhor ação a se realizar em cada situação do ambiente. Em contraste com

aprendizado supervisionado, em AR não se diz explicitamente ao aprendiz ou agente o que fazer ou qual a ação correta em uma dada situação. Ao invés disso, cada ação do agente é avaliada com um sinal numérico de recompensa, o qual dá ideia de qualidade imediata daquela ação. O próprio agente deve, por tentativa-e-erro, encontrar a ação que traz a maior quantidade de recompensas a longo prazo.

Dentre os métodos tradicionais de aprendizado por reforço, o clássico Q-learning [Watkins and Dayan 1992] é um dos pioneiros, e vários dos métodos mais bem sucedidos da atualidade usam seus princípios. No Q-learning, o agente mantém estimativas de valor (relacionado à soma das recompensas esperadas) para cada ação que possa executar em cada estado do ambiente. Ao interagir com o ambiente, o agente obtém uma amostra da recompensa real da ação que realizou no estado que estava. O Q-learning usa essa recompensa, além do valor do estado atingido, para atualizar suas estimativas. O Q-learning possui garantias teóricas de convergência de suas estimativas para os valores corretos [Watkins 1989]. Um agente treinado com Q-learning pode garantir o máximo possível de recompensa simplesmente selecionando a ação de maior valor em cada estado.

O Q-learning mantém as estimativas de valor das ações para cada estado em uma tabela. Se há muitos estados e/ou ações, tal representação não é viável. Uma maneira de resolver isso é usar uma função ao invés de uma tabela para as estimativas de valor. Tais funções podem receber representações contendo características dos estados e aplicar pesos, que podem ser aprendidos, para ponderar essas características [Sutton and Barto 2018, Parte II]. Em especial, a própria representação do estado pode ser aprendida, por exemplo, por uma rede neural profunda. Essa abordagem é colocada em prática no algoritmo Deep Q-Networks (DQN) [Mnih et al. 2015], o qual recebia pixels da tela e aprendeu a jogar jogos de Atari sem nenhum conhecimento prévio, obtendo desempenho sobrehumano em certos jogos.

Pode-se dizer que DQN inaugurou a "era do Aprendizado por Reforço Profundo", onde avanços substanciais continuam acontecendo. Tais avanços levaram métodos de aprendizado por reforço a obterem grande sucesso em jogos, desde jogos de tabuleiro [Silver et al. 2017b, Silver et al. 2017a] até video-games muito mais complexos que Atari [Berner et al. 2019, Vinyals et al. 2019], onde múltiplos jogadores devem responder rapidamente aos acontecimentos da tela enquanto traçam planos de longo prazo para vencer uma partida.

No entanto, a aplicação mais disruptiva de aprendizado por reforço foi em processamento de linguagem natural. Parte da metodologia de treino do ChatGPT consistiu em obter um modelo de recompensa através de *feedback* humano para textos gerados por um modelo pré-treinado e o refinamento do modelo gerador de textos para maximizar esta recompensa [Ouyang et al. 2022].

Para o leitor intressado, o livro de [Sutton and Barto 2018] é o principal livro-texto sobre aprendizado por reforço.

Parte II: Impactos da IA

Esta parte discute, de maneira não exaustiva, diversas áreas impactadas pela IA. Em cada área, são discutidas algumas aplicações, riscos e tendências. Inicialmente, discutimos "áreas meio", nas quais a IA tem relação com habilidades cognitivas humanas de visão (Seção 8.3) e linguagem (Seção 8.4). Avanços nas referidas áreas têm reflexo nas "áreas fim", nas quais a IA afeta diferentes aspectos da sociedade: Saúde (Seção 8.5), Indústria (Seção 8.6), Finanças (Seção 8.7) e Mobilidade Urbana (Seção 8.8).

8.3. Visão computacional e Processamento de imagens

Visão Computacional é um campo interdisciplinar que combina elementos de inteligência artificial, ótica e processamento de imagem. Trata-se de tornar as máquinas capazes de interpretar e extrair informações significativas a partir de imagens ou vídeos, permitindo a realização de tarefas como detecção de objetos, reconhecimento facial, análise de cenas, entre outras.

8.3.1. Aplicações

As áreas de visão computacional e processamento de imagens testemunharam enormes avanços nos últimos anos, em grande parte impulsionados por técnicas de aprendizado profundo. O aprendizado profundo, com sua capacidade de aprender automaticamente padrões complexos em grandes conjuntos de dados, revolucionou a maneira como abordamos as tarefas visuais. A sinergia entre visão computacional/processamento de imagens e aprendizado profundo levou a avanços significativos e abriu uma infinidade de aplicações práticas em vários setores e aplicações, algumas das quais brevemente listadas a seguir:

- Restauração e melhoria de imagens: algoritmos que integram processamento de imagens e aprendizado de máquina têm possibilitado a restauração visual de imagens degradadas, como remoção do ruído, aumento de resolução, correção de borramento por desfoco ou movimento, e correção de iluminação (sobretudo para imagens subexpostas).
- Detecção, Segmentação e Reconhecimento de Objetos: Algoritmos de detecção ou segmentação de objetos baseados em aprendizado profundo permitiram a identificação precisa e em tempo real de objetos em imagens e vídeos. Esta aplicação encontra uso prático em sistemas de vigilância, veículos autônomos e robótica. Por exemplo, carros autônomos usam a detecção de objetos para detectar pedestres, veículos e sinais de trânsito, e exploram segmentação para avaliar a área navegável.
- Reconhecimento facial e biometria: O reconhecimento facial é amplamente utilizado para fins de identificação e autenticação. Modelos de aprendizado profundo, como redes neurais convolucionais (CNNs), são capazes de identificar atributos faciais discriminatórios de cada indivíduo, permitindo o reconhecimento facial preciso mesmo em condições desafiadoras. Por exemplo, a biometria facial é empregada para autenticação de usuários em *smartphones*, sistemas de segurança e aplicativos de controle de acesso, agilizando processos e aumentando a segurança.

- Análise de Imagens Biomédicas: Conforme já discutido na Seção 8.5, o aprendizado profundo fez contribuições significativas para a análise de imagens biomédicas, auxiliando os profissionais de saúde a diagnosticar doenças com mais precisão e eficiência. Os modelos de aprendizado profundo podem detectar anomalias em raios-X, ressonâncias magnéticas e tomografias computadorizadas, auxiliando na detecção precoce de condições como câncer, doenças cardiovasculares e distúrbios neurológicos. Além disso, a visão computacional também facilitou a análise das lâminas histopatológicas, ajudando os patologistas a identificar e classificar as células cancerígenas.

- Realidade Aumentada (RA) e Realidade Virtual (RV): a visão computacional baseada em aprendizado profundo desempenha um papel crucial no desenvolvimento de aplicativos para RA e RV. Em particular, algoritmos de reconstrução tridimensional (3D) a partir de uma ou mais imagens podem ser usados para modelar ambientes reais, permitindo uma experiência imersiva com óculos de VR. Além disso, permitem rastrear e reconhecer objetos e cenas do mundo real, possibilitando a inserção de objetos sintéticos no ambiente do usuário (RA).

- Agronegócios e análise ambiental: a visão computacional baseada em aprendizado profundo transformou a agricultura com aplicativos como monitoramento de colheitas, detecção de doenças em vegetais e estimativa de rendimento. Drones equipados com câmeras e algoritmos de aprendizado profundo podem analisar vastas terras agrícolas, identificando áreas de preocupação e permitindo a aplicação precisa de fertilizantes e pesticidas. Além disso, o processamento de imagens aéreas e de satélite permitem o monitoramento de condições ambientais, como desmatamento, incêndios e enchentes.

- Varejo e *E-commerce*: A visão computacional também encontra lugar nos setores de varejo e comércio eletrônico. Os varejistas podem usar a visão computacional para rastrear o comportamento do cliente em suas lojas, analisar o tráfego de pedestres e otimizar os *layouts* das lojas para um melhor envolvimento do cliente. As plataformas de comércio eletrônico usam o reconhecimento de imagem para oferecer recomendações de produtos visualmente semelhantes, aprimorando a experiência de compra dos clientes.

- Automação Industrial: técnicas de visão computacional podem ser utilizadas em ambientes industriais, para a identificação de defeitos em produtos de maneira rápida e eficiente. Além disso, conjuntamente com a robótica, por exemplo, pode-se fazer com que robôs equipados com câmeras acessem ambientes de difícil acesso e realizem monitoramento continuado da infraestrutura.

8.3.2. Riscos

Apesar dos grandes avanços nos últimos anos, uma série de precauções devem ser tomadas antes do uso irrestrito de algoritmos de visão computacional baseados em aprendizado de máquina.

Um potencial problema se refere a *ataques*, nos quais uma imagem é manipulada para “enganar” um algoritmo de aprendizado de máquina. Por exemplo, Eykholt e

colegas [Eykholt et al. 2018] apresentaram uma estratégia para realizar *ataques físicos* focados no problema de detecção de sinais de trânsito (crucial para veículos autônomos). Como exemplo, mostraram que adesivos brancos e pretos colados a uma placa de trânsito mudam completamente o resultado de classificação de uma rede neural, apesar de manter o sinal completamente compreensível para o ser humano.

Um outro perigo potencial no uso de algoritmos de visão computacional baseados em aprendizado de máquina é a imprevisibilidade dos resultados em dados novos. Como a maioria das técnicas é supervisionada, são necessários dados de treinamento anotados. E como dados anotados são custosos de produzir e altamente dependentes do problema e aplicação, eles são disponibilizados em quantidade limitada. Por outro lado, aplicações de visão computacional envolvem dados nunca vistos pelas redes, que podem apresentar características distintas dos dados de treinamento, potencialmente causando degradação da qualidade. Essa variabilidade de características dos *datasets* é chamada de mudança de domínio (*domain shift*), e pode envolver diversos parâmetros (e.g., treinamento em dados capturados durante o dia e teste com dados capturados durante a noite). Como exemplo, Hasan et al. [Hasan et al. 2021] avaliaram o a capacidade de generalização de diversos algoritmos de detecção de pedestres, concluindo que a maioria apresenta resultados impressionantes em uma validação *intra-dataset*, mas com degradação acentuada em validações *cross-datasets*, mesmo quando o domínio alvo possui poucas diferenças visuais com relação ao domínio fonte.

Outra questão importante envolve os vieses, como os diversos problemas demonstrados por pesquisadores em sistemas de reconhecimento facial[3]. Notadamente, [Buolamwini and Gebru 2018] mostraram que algoritmos de reconhecimento facial discriminavam por raça e gênero.

8.3.3. Tendências

Por muito tempo, a grande maioria das redes neurais envolvendo imagens era baseada em camadas convolucionais. Embora elas tenha uma representação compacta em termos de número de parâmetros e estejam relacionadas com o funcionamento do sistema visual humano, o uso de convoluções assume um filtro com suporte espacial limitado. Assim, *pixels* muito distantes entre si podem não se relacionar em uma rede convolucional. Uma tendência crescente envolve o uso de camadas de atenção espacial, dentre as quais os *Transformers* são populares. O modelo Visual Transformer (ViT) [Dosovitskiy et al. 2020] estende o conceito de *Transformers*, originalmente desenvolvidos para textos, para o domínio de imagens.

Outra tendência atual é o uso integrado de dados visuais e textuais, dando origem às *Vision Language Models*. Como exemplo, o modelo CLIP (*Contrastive Language Image Pre-raining*) [Radford et al. 2021] usa uma base de 400 milhões de imagens pareadas com as respectivas descrições textuais, treinando os *embeddings* de texto e de imagens de tal maneira que eles sejam similares. Com esse tipo de abordagem, se pode fazer con-

[3]"Study finds gender and skin-type bias in commercial artificial-intelligence systems: Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women."MIT News, 11 de Fevereiro de 2018. https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212. Acesso em 28 de Agosto de 2023

sultas a imagens com dados de entrada textuais, e vice-versa. Com isso, diversos novos modelos e bases de dados têm sido propostas nos últimos anos. O *dataset* LAION-5B, por exemplo, fornece 5,85 bilhões de pares texto-imagens. Em particular, se mostrou que redes profundas treinadas com uma quantidade muito grande de dados (como CLIP) podem ser customizadas com poucos dados adicionais para tarefas específicas. Tais redes são chamadas de *Foundation Models*.

8.4. Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) é uma área multidisciplinar que combina ciência da computação e linguística com o objetivo de permitir que as máquinas realizem tarefas úteis com a linguagem humana. PLN geralmente envolve o processamento de grandes volumes de dados textuais ou de fala, que são chamados de *corpus* (ou *corpora* no plural). O PLN tem uma longa história e vem sendo estudado desde os anos 1950. A ideia de fazer com que os computadores consigam compreender a linguagem humana é tão antiga quanto a computação. Alguns dos exemplos mais icônicos vieram da ficção científica, como o HAL-9000 do filme “2001: Uma Odisseia no Espaço” dirigido por Stanley Kubrick (1968) e escrito por Kubrick e Artur C. Clarke, baseado em obra anterior deste intitulada "The Sentinel". O computador HAL-9000 exibia uma série de habilidades impressionantes envolvendo a compreensão e a geração de linguagem. Nos últimos anos, algumas dessas habilidades foram enfim atingidas por algoritmos reais. O imenso avanço em PLN se deve principalmente à evolução dos algoritmos de aprendizado profundo. Mais especificamente, o desenvolvimento da arquitetura de Transformers [Vaswani et al. 2017] usada nos grandes modelos de linguagem, do inglês *large language model* (LLM), como BERT [Devlin et al. 2019] e GPT [Brown et al. 2020], melhoraram sensivelmente os resultados em diversas tarefas de compreensão e de geração de texto. Pode-se dizer que as mudanças de maior impacto na imprensa e, notadamente disruptivas na computação nos últimos anos, vieram da área de PLN.

8.4.1. Aplicações

O PLN pode ser empregado em uma vasta gama de aplicações, tanto científicas como industriais. A seguir, fornecemos uma lista não exaustiva de aplicações.

Diversas tarefas de PLN podem ser modeladas como problemas de **classificação de textos**, por exemplo: análise de sentimento, filtragem de spam, identificação de discurso de ódio, atribuição de autoria, detecção de plágio, *etc*. Em todos esses casos, a entrada do algoritmo é uma sequência de tokens (*i.e.,* palavras) e a saída é a classe predita. Até meados dos anos 2000, os algoritmos mais empregados para classificação de textos eram os de aprendizado de máquina tradicional como Naïve Bayes, máquinas de vetores de suporte e árvores de decisão. A partir de 2014, as redes neurais profundas como o LSTM [Hochreiter and Schmidhuber 1997] passaram a predominar. De 2019 para cá, a hegemonia de modelos baseados em Transformers, principalmente o BERT [Devlin et al. 2019], é notável. O ganho de qualidade obtido nessas tarefas pode ser atribuído à ideia de trabalhar em duas etapas: o *pré-treinamento* e o *ajuste fino*. No pré-treinamento, o algoritmo analisa um corpus de grande volume (*i.e.,* contendo bilhões de palavras) a fim de aprender características da linguagem como relações entre palavras e coerência entre frases. O corpus usado nessa fase é apenas texto puro, que pode ser facilmente obtido a

partir da Web. Uma vez que o modelo tenha sido pré-treinado com grandes corpora, ele pode ser ajustado para desempenhar uma tarefa específica, como análise de sentimento, por exemplo. Nessa fase, o modelo precisa de um conjunto de dados rotulado com a classe esperada – o processo tradicional de aprendizado supervisionado. A vantagem aqui, é que o "conhecimento"que o modelo pré-treinado já possuía sobre a linguagem é aproveitado na tarefa específica e traz ganhos na qualidade das predições geradas.

A **sumarização de textos** tem por objetivo gerar uma versão reduzida do texto de entrada que contemple suas ideias principais. Diferentemente das tarefas de classificação em que a saída é uma só (*e.g.*, a classe predita), na sumarização a saída, assim como a entrada, é uma sequência de tokens. Algoritmos de sumarização podem ser empregados em diversos domínios como financeiro, jurídico, científico para reduzir a quantidade de texto que as pessoas precisam processar. As técnicas de sumarização dividem-se em *extrativas* e *abstrativas*. As técnicas extrativas usam técnicas estatísticas para selecionar as k frases mais significativas do texto original e as usam para compor o resumo. A desvantagem é que a coerência do texto gerado é prejudicada pois a conexão entre as frases pode ser perdida. Já as técnicas abstrativas usam LLMs para tentar simular o comportamento de um sumarizador humano combinando sentenças e utilizando paráfrases para gerar um texto mais fluido. A desvantagem aqui é que esses modelos podem "alucinar"e adicionar conteúdo no resumo que não tem base no texto original (veja mais sobre esse problema na Seção 8.4.2).

A **tradução automática** (TA) é outra aplicação importante de PLN que obteve melhorias significativas com a introdução dos algoritmos de aprendizado profundo. A tarefa consiste em transformar uma sequência de um idioma fonte para um idioma alvo. O exemplo mais conhecido é de tradutor automático é o GoogleTranslate[4]. Assim como a sumarização, tanto a entrada como a saída do algoritmo são sequências de tokens. A TA usa aprendizado supervisionado: durante o treinamento, o sistema processa um grande número de sentenças paralelas (as mesmas sentenças escritas no idioma fonte e no idioma alvo) e assim aprende o mapeamento entre idiomas. As arquiteturas mais usadas em sistemas de TA são LSTMs bidirecionais [Schuster and Paliwal 1997] e Transformers [Vaswani et al. 2017]. Além disso, vale ressaltar que esses sistemas trabalham em nível de subpalavras. Assim, eles conseguem aprender traduções para fragmentos, *e.g.*, o sufixo de gerúndio "*endo*"em português é normalmente traduzido para "*ing*"em inglês. A tradução automática é desafiadora por uma série de razões: (*i*) a ambiguidade que já é problemática em um idioma, fica ainda mais complexa quando adicionamos mais idiomas, (*ii*) a estrutura dos idiomas pode ser muito diferente e isso faz com que a ordem das palavras no idioma alvo possa ser muito diferente da ordem no idioma fonte, (*iii*) nem sempre há uma palavra correspondente no idioma alvo – pode não haver nenhuma ou podem haver várias traduções possíveis e (*iv*) vocabulários de domínios específicos (*e.g.*, medicina, computação, direito) dificilmente terão grandes volumes de sentenças paralelas para permitir a geração de mapeamentos precisos.

Os **chatbots**, também chamados de agentes conversacionais são programas que se comunicam com as pessoas usando linguagem natural. O exemplo mais marcante de

[4]`https://translate.google.com`

chatbot é o ChatGPT[5] que foi lançado em novembro de 2022 e em menos de dois meses já tinha mais de 100 milhões de usuários. O ChatGPT também é um modelo baseado em Transformers e suas capacidades vão além de conseguir manter diálogos, ele consegue escrever código de programas e compor músicas e poemas. Apesar de não ter representado um avanço científico (pois as tecnologias utilizadas já haviam sido justificadas), o impacto do ChatGPT foi gigantesco – poucas ferramentas geraram tanto interesse como essa. O mecanismo por trás dessas habilidades é a geração de texto que, assim como a geração de imagens, faz parte da IA generativa. A geração do texto usando LLMs é conhecida por geração autorregressiva ou geração causal. Ela consiste basicamente em escolher a próxima palavra a ser gerada condicionada às escolhas anteriores e à pergunta feita pelo usuário. Esse processo é conhecido como *next word prediction* (ou predição da próxima palavra). Essas escolhas dependem das estatísticas de ocorrência das palavras em grandes corpora de textos e em uma série de parâmetros que controlam a aleatoriedade do processo de geração. Os riscos desses sistemas são discutidos na Seção 8.4.2.

Até agora, esta seção abordou apenas texto. Contudo, PLN também trata de fala. A habilidade de **reconhecer e produzir fala** são muito relevantes e úteis em uma série de aplicações que interagem com os usuários por meio de voz como assistentes inteligentes (como a Siri da Apple e a Alexa da Amazon). A comunicação por meio de voz envolve o reconhecimento da fala (*Automatic Speech Recognition*) (ASR) que transforma de fala para texto e a conversão de texto para fala *Text to Speech* (TTS). ASR é uma tarefa bastante desafiadora pois precisa lidar com variações na forma de pronunciar as palavras (diferentes sotaques e velocidades de fala), ruídos de fundo e disfluências (sons como "hum"e "hã"). Tanto ASR como TTS atualmente são implementados utilizando LSTMs bidirecionais [Schuster and Paliwal 1997] ou Transformers [Vaswani et al. 2017]. Por ser mais difícil, ASR comumente precisa de mais dados de treinamento (*i.e.,* mais horas de áudio pareadas com o texto correspondente).

8.4.2. Riscos

Os principais riscos associados ao uso de aplicações de PLN advém, principalmente, de quatro problemas: o viés dos dados, as alucinações, o potencial para mau uso e o custo do treinamento de LLMs.

Os LLMs são treinados com grandes volumes de textos coletados a partir da web. Esses dados não passam por um processo de curadoria e podem conter diversos tipos de **viés** (racismo, sexismo, homofobia, xenofobia, *etc.*). O problema é que, ao gerar modelos a partir desses dados, os modelos passam a replicar esses vieses. [Papakyriakopoulos et al. 2020] observaram que até mesmo representações geradas a partir de textos da Wikipedia apresentam sexismo, homofobia e xenofobia. Também investigando vieses, mas na área de tradução automática, [Prates et al. 2020] mostraram que o Google Translate apresentava uma forte tendência de tradução para *defaults* masculinos em experimentos realizados a partir de uma lista abrangente de cargos do "Bureau of Labor Statistics"dos EUA. No artigo, traduções como “Ele/Ela é um Engenheiro” (onde “Engenheiro” é substituído por o cargo de interesse) em 12 idiomas diferentes de gênero neutro mostram que tradutor (que usa técnicas de IA) não consegue reproduzir uma distribuição real de traba-

[5] `https://chat.openai.com/`

lhadoras. O artigo mostra que o Google Translate produz padrões masculinos com muito mais frequência do que seria esperado apenas com base nos dados demográficos. Esses trabalhos indicam que é necessário o desenvolvimento de abordagens mais sofisticadas para a construção de sistemas que sigam princípios éticos, respeitando as diversidades populacionais, culturais, nacionais, de gênero, raça e muitas outras [Russell et al. 2015].

O segundo risco afeta os sistemas que geram texto de maneira autorregressiva: as **alucinações**. As alucinações referem-se a situações em que um modelo de linguagem gera texto que contém informações que não estão presentes nos dados de treinamento, ou seja, o modelo gera fatos falsos. Há vários casos que foram divulgados na imprensa e mídias sociais envolvendo desde erros mais inofensivos até a imputação de crimes a pessoas inocentes. A principal causa é a forma como esses modelos geram os textos: eles não têm nenhuma compreensão acerca da realidade que os textos descrevem. Pesquisadoras críticas dessa abordagem referem-se a esses modelos como "papagaios estocásticos"[Bender et al. 2021]. É importante ressaltar que ferramentas que apenas geram texto não substituem motores de busca (como o Google e Bing, por exemplo) pois elas não têm como apontar as fontes para as informações. Quando solicitadas, elas podem até mesmo criar referências falsas.

A qualidade dos textos gerados automaticamente pode ser útil em uma série de tarefas, mas por outro lado, abre possibilidades para o mau uso. Há relatos de advogados que usaram ferramentas como ChatGPT e Bard[6] para redigir processos, de candidatos a empregos que geraram currículos automaticamente contendo dados "inflados", de alunos que entregaram códigos de programa escritos pela ferramenta como sendo de sua autoria, de geração de notícias falsas, entre outros.

Por fim, com o aumento de ordens de grandeza no tamanho dos LLMs (*e.g.,* de 117 milhões de parâmetros do GPT2 para 175 bilhões no GPT3 – e um número desconhecido no GPT4) , o custo do treinamento desses modelos e o seu impacto ambiental também vêm sendo discutidos. Estimativas mencionam [Sharir et al. 2020] que o treinamento de um modelo com 1,5 bilhões de parâmetros possa chegar a US$ 1,6 bilhões.

8.4.3. Tendências

Sob a perspectiva acadêmica, obter contribuições de impacto em PLN está cada vez mais difícil pois as universidades com seus orçamentos reduzidos precisam competir com gigantes do mercado de tecnologia como a Microsoft e Google. Levando isso em consideração, um artigo recente de pesquisadores da Universidade de Michigan [Ignat et al. 2023] aponta algumas futuras direções de pesquisa. Dentre elas, destacamos o desenvolvimento de modelos multilíngues e para idiomas com poucos recursos, a incorporação de um raciocínio que tenha fundamentação no mundo real para reduzir o problema das alucinações, o investimento em interpretabilidade dos modelos para possibilitar que as predições sejam explicadas e a aplicação de PLN em domínios relevantes como a saúde e a educação.

A maturidade das técnicas de PLN e os bons resultados que vêm atingindo contribuem que elas sejam disseminadas e adotadas na indústria. A ampla disponibilidade de LLMs e modelos ajustados para as mais diversas tarefas facilita a sua implantação em

[6] `https://bard.google.com/`

sistemas, ferramentas e aplicativos que venham ser usado por um número cada vez maior de pessoas.

8.5. Saúde

A Saúde tem sido apontada desde cedo como uma das áreas de aplicação mais promissoras para a IA. Os primeiros exemplos de sucesso, ainda na década de 1970, tratavam-se de sistemas especialistas dependentes de conhecimento humano prévio e um conjunto de regras definidas para apoio à tomada de decisão clínica. Estes sistemas demonstraram utilidade para auxiliar na definição de diagnóstico ou na recomendação de tratamentos para pacientes, mas com um potencial muito limitado devido aos desafios de se representar um conhecimento complexo via regras e da incapacidade de extrapolar o conhecimento prévio a fim de aprimorar a tomada de decisão [Yu et al. 2018].

Desde então, impulsionada pelo aumento na disponibilidade de dados em saúde e pelo rápido progresso de algoritmos capazes de aprender padrões relevantes e acionáveis a partir de dados volumosos e complexos, a IA vem gradualmente revolucionando a área da Saúde. Aplicações inovadoras baseadas em IA, especialmente em aprendizado de máquina, estão provocando mudanças significativas na forma como abordamos a medicina e os cuidados de saúde, sejam individuais ou coletivos. Esta seção revisa os aspectos principais da intersecção entre IA e Saúde no que tange aplicações, riscos e tendências.

8.5.1. Aplicações

Embora praticamente todos os aspectos da prestação de cuidados de saúde sejam passíveis de uso e implementação de IA, quatro eixos se destacam nos esforços recentes, incluindo aqueles concentrados em países de baixa e média renda (LMICs, segundo sua sigla em inglês): (i) diagnóstico, (ii) avaliação do risco de morbidade ou mortalidade do paciente, (iii) previsão e vigilância de surtos de doenças e (iv) planejamento de políticas de saúde pública. [Schwalbe and Wahl 2020].

Sistemas para diagnóstico médico baseado em IA têm sido amplamente explorados em diversas áreas, mas alcançaram uma maturidade particular em especialidades como a radiologia, oftalmologia, patologia e dermatologia. Estes sistemas baseiam-se principalmente em dados de imagens médicas (e.g., ressonância magnética, tomografia computadorizada, fotografias de lesões ou de lâminas histopatológicas), demonstrando um desempenho diagnóstico via IA equivalente ao desempenho dos especialistas da saúde para casos de câncer de pele, câncer de mama, retinopatia diabética, doenças respiratórias, dentre outros [Liu et al. 2019]. Sinais biomédicos (e.g., eletrocardiograma e eletroencefalograma), exames laboratoriais, dados genéticos (e.g., mutações no DNA e expressão gênica) e informações de prontuários médicos eletrônicos também foram utilizados com sucesso nas mais diversas especialidades médicas, e a evolução da IA vem possibilitando que os médicos façam diagnósticos mais rápidos e precisos. A IA foi empregada, por exemplo, para estratificação de risco em pacientes com infarto do miocárdio por oclusão [Al-Zaiti et al. 2023], detecção precoce da doença de Alzheimer [Mahendran and PM 2022] e estimativa de risco de câncer de pulmão em 3 anos a partir de tomografia computadorizada e outras informações clínicas [Huang et al. 2019].

A IA também tem sido uma tecnologia fundamental para aprimorar a capacidade

de quantificar riscos de eventos desfavoráveis ou agravos relacionados à saúde de um paciente. Durante a pandemia da COVID-19, algoritmos de aprendizado de máquina foram amplamente aplicados para estimar quais pacientes infectados têm maior probabilidade de sofrer com uma doença mais severa ou vir a óbito pela COVID-19 ou suas complicações [Van der Schaar et al. 2021]. No estudo de [Phakhounthong et al. 2018], indicadores clínicos e laboratoriais foram utilizados para desenvolver um modelo baseado em IA para prever casos graves de dengue entre pacientes pediátricos durante a admissão. A avaliação de riscos propicia um melhor monitoramento do paciente e um tratamento mais efetivo através da antecipação de condutas clínicas, além de possibilitar uma melhor gestão de recursos hospitalares.

Os benefícios da IA também podem ser observados na análise de riscos de Saúde em nível coletivo ou populacional, contribuindo para o estudo da dinâmica de doenças e para uma melhor vigilância epidemiológica explorando uma grande variedade de dados, inclusive traços digitais (e.g., pesquisas na internet, atividades em redes sociais) [Brownstein et al. 2023]. [Jiang et al. 2018] utilizaram aprendizado de máquina para estimar a probabilidade de surto epidêmico de Zika em nível global, conseguindo melhor modelar a complexidade e não-linearidade da relação entre o risco de transmissão por Zika vírus e fatores climáticos, ambientais e sócio-econômicos. [Brownstein et al. 2023] apontam que a IA tornou-se grande aliada na vigilância de doenças infecciosas, viabilizando o desenvolvimento de sistemas de alerta precoce para surtos de doenças, a identificação de focos de surtos ou de patógenos causadores de doenças, o rastreamento de contato e a previsão eficaz do risco de transmissão. Assim, a IA possibilita que autoridades de saúde pública respondam adequadamente ao risco que se apresenta, por exemplo, alocando recursos ou suprimentos adequados diante da expectativa de aumento de casos de uma determinada doença em uma região. Adicionalmente, o uso da IA possui grande impacto no planejamento de políticas de saúde pública, ao permitir a elaboração de medidas mais eficazes para proteção e promoção da saúde, como o planejamento de campanhas de vacinação e do direcionamento de materiais de divulgação de prevenção e cuidados com saúde com base no perfil de risco pessoal e padrões comportamentais [Panch et al. 2019].

Por fim, a IA se estende à análise de dados moleculares e genômicos, desempenhando um papel fundamental nas pesquisas biomédicas e possibilitando expandir nosso conhecimento sobre o funcionamento das doenças. A IA tem sido uma das principais propulsoras da medicina de precisão, especialmente na área da oncologia, revelando assinaturas moleculares associadas a subtipos ou estágios tumorais [Marczyk et al. 2023], e identificando novos biomarcadores [Colombelli et al. 2022], incluindo aqueles úteis para detecção não invasiva de câncer e avaliação de prognóstico [Xu et al. 2019]. Adicionalmente, avanços recentes como o AlphaFold [Jumper et al. 2021], que se utiliza de aprendizado profundo para prever com alta precisão as estruturas tridimensionais das proteínas a partir de sua sequência de aminoácidos, permitem facilmente avaliar o impacto funcional de variantes genéticas e acelerar a descoberta de novas drogas.

8.5.2. Riscos

Apesar do notável aumento nas pesquisas relacionadas às aplicações da IA na área da Saúde, é importante destacar que apenas um conjunto limitado destas soluções foi efetivamente implementado na prática clínica [Rajpurkar et al. 2022]. A Food and Drug

Administration (FDA), agência reguladora vinculada ao Departamento de Saúde e Serviços Humanos dos Estados Unidos, tem desempenhado um papel ativo na revisão e autorização para comercialização de um número crescente de dispositivos médicos que incorporam IA. No entanto, até a última atualização em outubro de 2022[7], constatou-se a aprovação de apenas 521 dispositivos médicos pelo FDA. Dentre esses dispositivos, 56,23% são voltados para aplicações em Radiologia e 10,94% na Cardiologia. A disparidade evidente entre a extensa quantidade de pesquisas científicas conduzidas nesse domínio e o número limitado de soluções práticas adotadas reflete uma série de desafios que permeiam a integração da IA na prática médica.

Embora diversos fatores possam contribuir para esse cenário, existe um consenso na comunidade acadêmica de que a falta de validação dos modelos baseados em IA por meio de dados externos constitui um dos principais fatores que inibem a aplicação efetiva do conhecimento científico adquirido [Liu et al. 2019]. Esta validação deveria ser feita com dados prospectivamente coletados a partir do mundo real para este propósito específico, e seguindo uma metodologia criteriosa de avaliação, como aquelas adotadas em ensaios clínicos randomizados. Modelos de IA podem falhar na generalização para novos tipos de dados nos quais não foram treinados. Alguns trabalhos já demonstram que a capacidade preditiva de um modelo é impactada negativamente quando o modelo é aplicado a uma população de pacientes diferente dos seus dados de treinamento [Wong et al. 2021]. Isto se deve à ampla heterogeneidade dos dados neste domínio devido a diferenças existentes nas práticas hospitalares e nos dados demográficos dos pacientes entre diferentes hospitais.

Outro ponto crítico é que o treinamento de modelos de IA em conjuntos de dados com pouca representatividade de grupos marginalizados ou com variações injustificadas para determinados grupos resulta em sistemas tendenciosos que apresentam baixo desempenho preditivo nesses grupos. Assim, sem o controle adequado, o uso da IA introduz o risco de perpetuar vieses ocultos nos dados e reforçar preconceitos e desigualdades sociais existentes. Por exemplo, um viés racial foi detectado em um algoritmo de avaliação de risco clínico utilizado nos Estados Unidos, que atribuía menor risco a pacientes negros em comparação com pacientes brancos igualmente doentes por utilizar custos de saúde como um proxy para as necessidades de saúde [Obermeyer et al. 2019]. Em outro estudo, um viés étnico foi identificado em escores de risco poligênico utilizados para estimar o risco de um indivíduo desenvolver doenças como câncer com base em fatores genéticos, possuindo acurácia muito superior em indivíduos de ascendência Europeia do que para outras ancestralidades em razão da coleta desequilibrada de dados genéticos e genômicos entre continentes [Martin et al. 2019].

Estes riscos são exacerbados na impossibilidade de explicar a tomada de decisão pelos modelos e avaliar até que ponto a mesma reflete as abordagens humanas especializadas e não fere princípios éticos fundamentais. A Organização Mundial da Saúde (OMS) [World Health Organization 2021] chama atenção, ainda, para os vieses derivados de exclusão digital. Em alguns LMICs, mulheres têm menos acesso a telefone celular

[7] `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices`. Acesso em 14 de Agosto de 2023

ou internet móvel do que homens, contribuindo com menos dados para treinamento de modelos de IA e sendo menos propensas a se beneficiar do uso desta tecnologia [World Health Organization 2021]. A fim de gerar modelos baseados em IA que possam promover equidade em Saúde, é imprescindível garantir disponibilidade e qualidade de dados, com diversidade em relação a contextos sociais, culturais e econômicos. Por fim, é inevitável apontar o risco de violação da privacidade por se tratar de dados sensíveis, e o risco de uso indevido de dados pessoais, visto que muitas vezes os modelos são treinados com base de dados retrospectivas, originalmente coletadas para outros propósitos de pesquisa.

8.5.3. Tendências

A Organização Mundial da Saúde (OMS) [World Health Organization 2021] reconhece o enorme potencial da IA para promover melhorias na medicina e alavancar a equidade dos cuidados em saúde. Para que este potencial se concretize, avanços ainda se fazem necessários em diversas frentes, sendo algumas mais críticas para o domínio da saúde.

Técnicas para mitigar vieses, sejam estes oriundos dos próprios dados ou resultantes do processo de treinamento dos modelos, são primordiais para evitar que se perpetuem desigualdades sociais existentes ou que se introduzam comportamentos tendenciosos nos modelos que possam produzir resultados discriminatórios contra determinados grupos. Entretanto, garantir maior equidade através do uso dos modelos também requer uma capacidade mais apurada de explicar as predições realizadas pelos mesmos a fim de identificar erros sistemáticos indesejáveis. Neste sentido, pesquisas em torno da explicabilidade de modelos são essenciais a fim de expandir a capacidade de encontrar fatores relevantes para as predições realizadas com base não somente em associações, mas em relações causais entre as variáveis de entrada e o resultado do modelo. Uma explicabilidade baseada em causalidade tornaria a interação especialistas-IA muito mais efetiva para a investigação da tomada de decisão feita pelos modelos, resultando em maior confiança na implementação prática destes modelos.

Por fim, salienta-se que a tomada de decisão em um ambiente clínico é inerentemente baseada em múltiplas evidências, sendo portanto crucial ampliar a capacidade dos algoritmos de aprenderem a partir de dados multimodais. Desta forma, modelos multimodais de IA visam possibilitar o uso de todas as fontes de dados normalmente disponíveis aos médicos ou que possam enriquecer a definição de um diagnostico, como dados clínicos e laboratoriais, exames de imagens, testes genéticos, determinantes sociais, fatores ambientais ou comportamentais, informações coletadas por *wearables*, dentre outros. Estas direções de pesquisa estão entre os principais pontos de acesso para fornecer cuidados ao paciente mais oportunos, precisos e justos com auxílio da IA.

8.6. Indústria

Da mesma forma que a eletricidade transformou drasticamente a indústria na segunda revolução industrial, a IA tem sido apontada como uma das promotoras de grandes transformações na indústria atualmente. Nos últimos anos, temos testemunhado uma crescente adoção de abordagens de IA nos mais diferentes setores da indústria, tais como energia [Pivetta et al. 2023, Rahmanifard and Plaksina 2019], manufatura [Li et al. 2017], indústria química [Baum et al. 2021], agricultura industrial [Benos et al. 2021] , etc.

Atualmente, a IA é apontada como um fator viabilizador com papel crucial na chamada *indústria 4.0* (quarta revolução industrial), que tem como característica fundamental o foco no desenvolvimento de *indústrias inteligentes*. Este cenário surge graças ao desenvolvimento e integração da IA com diferentes tecnologias, tais como redes de dados, internet das coisas, computação em nuvem, automação de processos físicos, sistemas ciber-físicos, etc.

Em um cenário típico de uma indústria inteligente alinhada à industria 4.0, a fábrica é constituída por coleções de sistemas ciber-físicos, que estabelecem uma interação profunda entre processos físicos e processos computacionais. Neste contexto, processos computacionais distribuídos controlam elementos físicos e sensores realizam continuamente o monitoramento dos processos e componentes físicos, retroalimentando os processos computacionais. Neste contexto, a IA desempenha um papel crucial na tomada de decisão que controla estes processos continuamente [dos Anjos et al. 2023].

A seguir, serão discutidas algumas aplicações de IA na indústria, bem como os riscos e as tendências associadas ao uso de IA neste contexto.

8.6.1. Aplicações

A IA vem sendo aplicada nos mais diversos setores industriais, das mais diferentes formas, incluindo otimização e automação de processos produtivos, previsão de demandas e de produção, identificação de perfil de clientes, desenvolvimento de produtos com IA, automação de atendimento ao cliente, desenvolvimento de novos produtos, etc. Nestes contextos, a aplicação de técnicas de IA visa aumentar a produtividade, reduzir custos, tornar as linhas de produção mais seguras, etc.

Uma das aplicações mais notórias da IA na indústria diz respeito ao uso destas tecnologias para automação de processos produtivos [Ribeiro et al. 2021, Fragapane et al. 2022]. A automação, neste caso, envolve principalmente a utilização de robôs, ou sensores e atuadores distribuídos ao longo das linhas de produção. Neste cenário, sistemas de IA utilizam dados de sensores para tomar decisões e controlar os atuadores ao longo da linha de produção. Nestes cenários, o uso da IA promove o aumento da produtividade e o desenvolvimento de linhas de produção mais flexíveis.

Técnicas de IA também vêm sendo largamente utilizadas na indústria para detectar anomalias em comportamentos de sistemas, de processos produtivos, etc [Stojanovic et al. 2016, Zipfel et al. 2023]. Anomalias, neste cenários, são padrões de comportamento diferentes do comportamento esperado [Martí et al. 2015]. Nestes contextos, em geral, são aplicadas técnicas de aprendizado de máquina para treinar algoritmos que identifiquem estados normais e anômalos dos sistemas e processos de interesse. Estes algoritmos costumam ser treinados a partir de dados de sensores que caracterizam os estados dos processos e sistemas ao longo do tempo. Esta abordagem é utilizada, por exemplo, para detecção em tempo real de possíveis vazamentos em oleodutos [Aljameel et al. 2022]. Em alguns casos, anomalias podem ser detectadas do modo visual também [Roth et al. 2022], em cenários em que as anomalias não são bem representadas por medidas de sensores convencionais (como medidas de pressão e temperatura, etc), mas se tornam aparentes através da inspeção visual. Estas abordagens são muito comuns, por exemplo, para detectar defeitos em produtos em linhas de produção [Birlutiu et al. 2017],

permitindo a remoção do produto defeituoso do processo para eventuais correções dos defeitos. Nestas abordagens, técnicas de aprendizado de máquina podem ser utilizadas para aprender padrões que caracterizam produtos com e sem defeitos a partir de grandes conjuntos de imagens previamente rotuladas.

Além de aperfeiçoar os processos produtivos, tecnologias de IA também vêm sendo utilizadas na indústria para a previsão de demandas e para o gerenciamento da cadeia de suprimentos necessários para suprir estas demandas [Zhu et al. 2021, Toorajipour et al. 2021], incluindo a previsão de oferta de insumos e seleção de fornecedores. Muitas das aplicações nesta área vêm utilizando técnicas de aprendizado de máquina capazes de lidar com dados em séries temporais.

No contexto da indústria 4.0, os sistemas produtivos tendem a ser altamente sensorizados, de modo que uma grande quantidade de medidas são continuamente adquiridas dos equipamentos ao longo do tempo. Neste cenário, estes dados podem fornecer valiosos *insights* sobre o estado dos equipamentos. Estes fatores vêm permitindo o desenvolvimento de técnicas de *manutenção preditiva* [Paolanti et al. 2018, Paolanti et al. 2018, Serradilla et al. 2022, Dalzochio et al. 2020] baseadas em técnicas de aprendizado de máquina. Abordagens de manutenção preditiva visam monitorar o estado dos equipamentos com o intuito de prever eventuais momentos de falha antes que elas ocorram, permitindo a redução de custos oriundos de paradas não programadas na produção, ou ainda evitando falhas que podem comprometer drasticamente as plantas de produção.

Nos últimos anos, a IA vem sendo utilizada até mesmo no processo de *design* de novos produtos [Aphirakmethawong et al. 2022]. Aplicações típicas de IA em design de produtos vêm utilizando as mais diversas abordagens de IA, incluindo desde algoritmos genéticos [Kielarova and Pradujphongphet 2023] a aprendizado de máquina [Zhang et al. 2019, Fournier-Viger et al. 2021, Hamolia and Melnyk 2021]. Cabe destacar que nos últimos anos técnicas de IA generativa vêm demonstrando capacidades impressionantes em tarefas de design [Grisoni et al. 2021]. Modelos de IA generativa, como ChatGPT e Dall-E, são capazes de aprender padrões a partir de grandes massas de dados (imagens, textos, etc) e gerar saídas que reproduzem esses padrões de forma verossímil. Algumas aplicações representativas de IA em design de produtos incluem o projeto de circuitos integrados [Gubbi et al. 2022, Wang and Luo 2019, Hamolia and Melnyk 2021], desenvolvimento de novas drogas [Grisoni et al. 2021], desenvolvimento de peças de vestuário na indústria da moda [Liang et al. 2020, Giri et al. 2019], desenvolvimento de produtos na indústria alimentícia [Zhang et al. 2019], etc.

É importante salientar que as aplicações industriais de técnicas de IA são vastas, abrangendo muitos setores e muitas tarefas diferentes, de modo que nesta seção são discutidos apenas alguns exemplos.

8.6.2. Riscos

A aplicação da IA na indústria herda boa parte dos riscos da IA aplicada em contextos gerais. Um destes riscos, e que pode impactar aplicações industriais de diversas formas, é o da falta de generalização de modelos de aprendizado de máquina. Em caso de falha na generalização destes modelos, sistemas de IA podem cometer erros em casos em que precisam lidar com situações muito diferentes das representadas nos dados de treinamento

ou com dados capturados por sensores com características técnicas diferentes dos sensores que coletaram os dados de treinamento. Em contextos industriais, erros no processo de decisão acarretados por modelos que não generalizaram adequadamente podem causar diversos impactos negativos. Por exemplo, em casos em que o ambiente industrial possui atuadores controlados por modelos sem a devida generalização, falhas no processo de decisão podem disparar ações (como movimentos de braços robóticos) que podem eventualmente ferir seres humanos que também atuam no ambiente industrial [Franklin et al. 2020]. Outros exemplos do impacto negativo da falta de generalização incluem detectar incorretamente defeitos em produtos, o que pode fazer com que produtos defeituosos sejam mantidos ou produtos sem defeitos sejam removidos das linhas de produção.

Apesar da extensa pesquisa na área de aprendizado de máquina visando encontrar maneiras de mitigar o problema da generalização, ainda existem diversos desafios relacionados à própria identificação adequada do domínio de validade dos modelos de aprendizado de máquina. Ou seja, dado um modelo de aprendizado de máquina treinado em um certo conjunto de dados, não é trivial determinar quais são os conjuntos de situações em que ele funciona adequadamente ou não. Esta dificuldade pode tornar modelos de aprendizado de máquina suscetíveis aos chamados ataques adversários [Narodytska and Kasiviswanathan 2017], em que alguém mal intencionado pode alterar sutilmente os dados de entrada (de um modo imperceptível para seres humanos) de certos modelos com o intuito de perverter o comportamento esperado. Estas dificuldades estão em grande parte associadas à dificuldade de se explicar de forma significativa o que de fato foi aprendido pelo modelo. Essa dificuldade associada à explicabilidade de modelos de aprendizado de máquina vem sendo apontada como um risco pela indústria em geral.

Além disso, atualmente há uma grande discussão a respeito das consequências da aplicação da IA no mercado de trabalho [Agrawal et al. 2019]. A utilização da IA na indústria, em geral, promove um aumento da automatização dos mais diferentes processos. No passado, os processos de automatização atingiram principalmente os aspectos físicos dos processos industriais. Mas com aplicações de tecnologias de IA estamos testemunhando também o impacto em aspectos intelectuais do trabalho. Esta tendência gera ainda mais impactos na oferta de empregos, diminuindo a oferta de certos postos, mas eventualmente proporcionando o surgimento de novas profissões.

8.6.3. Tendências

Uma pergunta-chave, tendo em vista a ubiquidade da IA, é como identificar tendências relevantes para negócios? Um análise ampla de tendências na área de IA pode ser realizada de diversas formas. Muitas vezes, a abordagem acadêmica utilizada é da identificação de áreas de classificação de artigos em publicações. No entanto, esta é uma abordagem obviamente limitada às bases de consulta e a preferências das conferências e revistas no que se refere a áreas de pesquisas. O atual impacto da IA - ressalte-se - surge a partir de uma subárea que era pouco valorizada na academia por um período de mais de uma década: redes neurais artificiais. Entre meados da década de 1990 até 2006, quando Geoffrey Hinton e seus alunos publicaram o primeiro artigo no qual os autores se referem

a redes neurais profundas[8] [Hinton et al. 2006], poucos autores consideravam o aprendizado conexionista como sendo uma grande tendência futura em IA. Assim, pensar em tendências, como diria Niels Bohr, é prever o futuro - e não há algo mais difícil do que prever o futuro em ciência.[9]

Consultorias especializadas em tecnologia, como Gartner, IDC, McKinsey e diversas outras analisam e identificam periodicamente diversas áreas da computação que terão impacto ao longo do tempo, bem como sua maturidade [10]. Do ponto de vista de mercado, tais estudos têm grande relevância, pois orientam profissionais e gestores. Na última década, outros relatórios sobre análise e tendências em IA têm sido publicados por centros de pesquisa, reunindo parcerias entre a academia e as empresas. Entre eles destacamos o AI Index Report, produzido sob a coordenação do Human-Centered AI Institute da Universidade de Stanford[11]. Este relatório, publicado anualmente desde 2017, destaca as tendências em pesquisa e desenvolvimente (através de análise de publicações), performance técnica (onde se analisam os progressos tecnológicos e seus impactos), ética em IA (equidade, viéses e suas implicações), impacto econômico (utlização da IA em negócios, investimentos públicos e privados), educação (nas escolas e universidades), políticas e governança (estratégias nacionais e multilaterais de governos), diversidade (notadamente na academia e as iniciativas para seu incremento) e opinião pública (análise da percepção pública sobre o impacto da IA). Uma observação relevante do AI Index Report[12] é que até 2014, a maior parte dos sistemas de aprendizado de máquina eram produzidos pela academia. Desde então, as empresas passaram a dominar a produção destas tecnologias. Os dados do relatório indicam que produzir sistemas de aprendizado de máquina de estado-da-arte requer grandes volumes de dados, poder computacional e recursos financeiros não disponíveis às universidades. Isto pode sugerir que, assim como demais tecnologias do passado, a partir do momento em que a viabilidade técnica e o alto potencial econômico de uma tecnologia são demonstradas, os investimentos neste setor tendem e se consolidar nas empresas.

8.7. Mercado de Capitais e Finanças

Finanças tem dois aspectos interessantes, sendo ao mesmo tempo uma arte e uma ciência. Então, pode-se entender finanças como a arte e a ciência da gestão de ativos financeiros. Mas as pessoas frequentemente se deparam a questão recorrente a seguir: *porque eu deveria me interessar pelo que acontece no mercado de capitais?*. Para abordar esta questão seria interessante discutir o que acontece no mercado de capitais.

O mercado de capitais é para onde os governos recorrem para fechar suas contas (geralmente pedindo empréstimos pela venda de pequenos 'pedaços' da dívida do governo, ex: títulos de dívida como os conhecidos 'títulos do tesouro'). A razão principal

[8]Deep neural networks, no caso do artigo [Hinton et al. 2006] se referem a "deep belief networks", uma arquitetura de redes neurais para aprendizado.

[9]"Prediction is very difficult, especially if it's about the future."Frase atribuída a Niels Bohr e, também, a Yogi Berra.

[10]`https://www.gartner.com` e `https://www.idc.com/`

[11]`https://aiindex.stanford.edu/report/`

[12]AI Index Report 2023, Capítulo 1, Página 50. `https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf`

para o governo ir ao mercado de capitais é honrar seus compromissos, tais como pagar benefícios sociais (ex: aposentadorias, seguro-desemprego, etc.), pagar outras contas obrigatórias (ex: saúde, educação, etc.), ou desenvolver seus projetos (ex: casa própria, saneamento básico, etc.). As empresas públicas e privadas também recorrem ao mercado de capitais para pedir empréstimos e cumprir com suas obrigações (ex: pagar dívidas vencendo em prazos curtos), desenvolver seus projetos (ex: investir em serviços licitados tais a expansão da rede de saneamento básico, do sistema de geração e distribuição de energia, etc.). As pessoas físicas também recorrem ao mercado de capitais para obter recursos e atingir diversos objetivos (ex: projetos futuros, casa própria, aposentadoria, etc.).

Como o mercado de capitais mobiliza a poupança, gere riscos, aloca eficientemente recursos e promove o aumento da disciplina corporativa, toda a sociedade é beneficiada. Ao aplicar sua poupança em capital produtivo, os investidores (individuais ou institucionais) causam movimentos de capitais e buscam uma alocação eficiente e com menor custo. Isso aumenta a liquidez da economia e os prazos dos investimentos. Para fazer essa alocação de capitais (recursos), os participantes do mercado exigem qualidade na governança corporativa e o compartilhamento de informação por parte das empresas que captam estes recursos, levando a maior disciplina e transparência, o que impacta na produtividade e no retorno sobre o investimento realizado. O resultado final, em um nível macro-econômico, é mais emprego, renda, investimento e crescimento econômico, o que impacta positivamente os principais indicadores socioeconômicos do país. Esse ciclo virtuoso apontado acima também traz resultados indiretos, tais como a viabilização e o desenvolvimento de mais projetos, a expansão da produção e da criação de riqueza, a criação de mais empregos, e o aumento da renda do cidadão. Portanto, desenvolvendo o mercado de capitais, promove-se o desenvolvimento socioeconômicos do país.

Então, parece natural que alguém tenha interesse pelo que acontece no mercado de capitais e no mundo das finanças, mas para identificar a relação entre a computação, o mercado de capitais e o mundo das finanças, precisamos observar mais detalhadamente o que acontece no dia-a-dia do mercado de capitais.

Qualquer transação no mercado de capitais envolve um acordo entre duas partes: (a) tomador de capital e (b) provedor de capital. Geralmente, quem provê capital o faz esperando algum retorno (ex: um valor referente ao aluguel do capital via juros, um lucro para remunerar o capital investido em uma transação, etc.). Por outro lado, quem toma capital também o faz esperando algum retorno (ex: atingir o objetivo de adquirir um bem de capital ou ativo real, desenvolver um projeto, etc.). Como há muitas partes interagindo direta ou indiretamente no mercado de capitais, com propósitos muito diferentes, existe uma grande variedade de instrumentos disponíveis para atender aos diferentes interesses das partes, possibilitado que elas interajam e atinjam seus objetivos através do uso de instrumentos específicos. Estas interações podem ocorrer em diferentes ambientes (ex: há ambientes em que existe livre negociação de instrumentos entre partes, como nas bolsas de valores, e existem ambientes onde a negociação ocorre dentro de restrições, como no caso das interações entre as empresas e seus clientes). Esta diversidade de tipos de instrumentos transacionados entre partes e o grande volume de transações tendem a tornar o dia-a-dia do mercado de capitais complexa. Cada instrumento transacionado entre partes provê informações sobre o mercado, seus segmentos, partes envolvidas e sobre a própria economia, em diferentes níveis (ex: micro ou macro-econômico, local, regional, nacional

ou mesmo internacional). Devido a grande complexidade das operações realizadas, da necessidade de rastrear e armazenar o enorme volume de informações gerado, o mercado de capitais migrou quase em sua totalidade para o ambiente digital e as transações são computadorizadas.

8.7.1. Aplicações

Hoje, a IA participa da automação de tarefas rotineiras, prove acessibilidade a serviços, e capacidade de aprendizado para aperfeiçoar tarefas rotineiras nos mercados, de forma exata, eficiente e rápida. Alguns dos temas abordados com o auxílio da IA no dia-a-dia do mercado de capitais e em finanças são: 1) análise ou inferência de sentimento dos participantes do mercado com base em textos, ou mesmo de mídias sociais; 2) detecção de transações suspeitas ou fraudulentas, ameaças e/ou crimes financeiros, ou ainda ameaças cibernéticas; 3) identificação de riscos e vantagens potenciais de ativos transacionados nos mercados; 4) avaliação e recomendação de instrumentos financeiros (ex: produtos e serviços) para potenciais interessados(as), com base nas suas preferências e objetivos; 5) processamento de dados estruturados e não estruturados, tais como documentos, para extrair dados relevantes e alimentar os processos de análise, previsão e recomendação (ex: descoberta de oportunidades de investimento); 6) uso de estimativas de risco, dados de transações e complementares para prever com razoabilidade resultados futuros; 7) sintetize de informações relevantes para a tomada de decisão usando IA generativa.

Por exemplo, já é comum alguém 'falar' com um robot ao tentar abrir uma conta bancária pela internet, ou mesmo ser entrevistado(a) por um robot para direcionar uma chamada telefônica a uma instituição financeira. O Business Insider[13] estima que as aplicações de IA vão economizar para as instituições financeiras nos EUA, em 2023, cerca de USD 447 bilhões. A maioria dos bancos (cerca de 80%) avaliam como positivo o impacto da IA neste segmento da indústria, e pretendem acelerar a migração das suas operações físicas para o ambiente digital (ex: um dos maiores investimentos do Banco Mercantil do Brasil em 2023 é focado no aumento da acessibilidade digital). Mani Nagasundaram (Senior VP, Global Financial Services, HCL Technologies) afirmou recentemente em um artigo na AI News[14] que a IA tende a liberar pessoal de tarefas rotineiras, e ao mesmo tempo melhorar a qualidade e a segurança do acesso a serviços financeiros, além de contribuir para trazer inovação ao ambiente corporativo. Já a Forbes[15] sugere em um artigo recente que 70% das empresas do setor financeiro já usam IA para prever eventos que possam afetar o seu fluxo de caixa (antecipando a necessidade de caixa para as operações diárias), ajustar os limites de crédito dos clientes e detectar fraudes no uso dos serviços (ex: cartões de crédito). Também, de acordo com a Forbes[16], IA tem sido usada para identificar as tendências mais recentes dos mercados, e avaliar os perfis das carteiras de

[13] https://www.insiderintelligence.com/insights/ai-in-finance/

[14] https://www.artificialintelligence-news.com/2020/12/15/from-experimentation-to-implementation-how-ai-is-proving-its-worth-in-financial-services/

[15] https://www.forbes.com/sites/louiscolumbus/2020/10/31/the-state-of-ai-adoption-in-financial-services/?sh=711a8c4e2aac

[16] https://www.forbes.com/sites/jayadkisson/2019/01/23/artificial-intelligence-will-replace-your-financial-adviser-and-thats-a-good-thing/?sh=1a02118e6b40

investimento disponíveis e dos clientes, para então sugerir os quais investimentos seriam adequados para cada perfil de cliente. Como IA é usada frequentemente para analisar padrões em grandes conjuntos de dados, naturalmente esta habilidade da IA tem sido usada em negociações nos mercados abertos (ex: bolsas). Pois os métodos computacionais baseados em IA podem analisar dados mais rápido e com maior exatidão que os humanos, além de poderem aprender a serem mais eficientes e otimizar as negociações nestes mercados (ex: algoritmos inteligentes já são usados para achar interessados em fixar a taxa de conversão Dolar-Real de um contrato de exportação, e os interessados em contratar esta taxa de conversão Dolar-Real na data futura desejada; ou ainda, algoritmos inteligentes já são usados para identificar tendências nos mercados e sugerir transações de ativos financeiros que estejam alinhadas com estas tendências).

8.7.2. Riscos

Não se pode ignorar que existem desafios éticos e riscos a serem mitigados para que o uso da IA nos mercados e em finanças seja efetivo, especialmente no que se refere a proteção de informações sensíveis e financeiras dos participantes, e a proteção dos participantes dos riscos que o uso das informações providas por algoritmos podem trazer. O Fintech Times [17] aponta três temas sensíveis que merecem atenção ao introduzir recursos de IA no setor financeiro e nos mercados em geral:

- *Ausência de Viés*: Antever que podem ocorrer falhas no projeto de algoritmos, e consequências indesejadas. Por exemplo, um algoritmo falho poderia adquirir um comportamento predatório ao considerar a necessidade da instituição financeira de otimizar sua lucratividade nas operações. Um algoritmo falho poderia se tornar predatório errando na estimativa da capacidade de um cliente de se endividar e de assumir riscos, e então oferecer ativos muito arriscados para este cliente que não teria condições de correr tais riscos nem de assumir as dívidas associadas a eles;

- *Responsabilização e Regulação*: É importante: a) regulamentar quais fontes de informações podem ser usadas e o uso correto destas informações, e b) definir antecipadamente de quem seria a responsabilidade se houverem consequências indesejadas de possíveis erros gerados por algoritmos, e como lidar com estas consequências (ex: informações incorretas ou inexatas podem induzir a erros na tomada de decisão e/ou em transações);

- *Transparência*: O que levou o algoritmo a tomar uma decisão, ou executar uma ação, deveria ser conhecido e rastreável.

Também é importante chamar a atenção para o fato de que *algoritmos podem ser usados como armas*. Portanto, o uso de algoritmos para propósitos antiéticos, tais como roubo de informações sensíveis de clientes, deve ser evitado e responsabilizado.

[17] `https://thefintechtimes.com/the-ethics-of-ai-ai-in-the-financial-services-sector-grand-opportunities-and-great-challenges/`

8.7.3. Tendências

Ao consolidar tarefas e analisar dados de forma mais rápida e exata que os humanos, espera-se que o uso intensivo de IA economize mais de USD 1 trilhão para os bancos e instituições financeiras nos EUA até 2030[18]. McKinsey Co.[19] estima que o sistema financeiro deverá ser transformado pelas mudanças tecnológicas, e as instituições financeiras precisarão aumentar seus investimentos em tecnologia da informação e IA para atingir altos níveis de digitalização, com qualidade. Será uma questão de sobrevivência, pois McKinsey Co. também estima que mais de 78% dos clientes jovens não iriam na agência física de uma instituição financeira se tivessem uma alternativa.

8.8. Mobilidade Urbana

A agenda em torno de cidades inteligentes tem como um dos focos a mobilidade urbana inteligente (uso racional dos diversos meios de transporte, integrando-os e adaptando-os à demanda). Existem diversas possibilidades em relação ao uso de IA em geral – e aprendizado de máquina em particular – em tal agenda. O restante desta seção joga luz em alguns aspectos a respeito de como a IA vem contribuindo, e como seu papel se torna cada vez mais decisivo. Desta forma, um verdadeiro trânsito inteligente resultará de indivíduos, semáforos e veículos conectados e trabalhando em conjunto. Nesta visão, semáforos inteligentes são alimentados com informação a respeito do estado da rede de tráfego, sobre os sobre semáforos vizinhos, eventos imprevistos, e outras informações.

Uma explicação mais detalhada sobre sistemas de transporte e simulação de tráfego pode ser encontrada em [Bazzan and Klügl 2013, Bazzan 2021]. A seguir, serão abordados dois problemas centrais, os quais motivam diversas aplicações de IA. O primeiro se dá pelo lado da demanda (como se deslocar de A até B de maneira eficiente), enquanto que o segundo se refere ao lado da oferta (controle e gerenciamento de tráfego). Devido à limitação de espaço, nos concentramos no tráfego veicular urbano.

8.8.1. Aplicações

Em relação à oferta, quando se fala em mobilidade inteligente, as pessoas em geral pensam em semáforos inteligentes. Algoritmos e técnicas de controle semafórico existem há várias décadas e derivam principalmente de técnicas de pesquisa operacional e da área de controle. Mais recentemente, técnicas de IA e aprendizado de máquina têm sido empregadas, em especial aqueles que se baseiam em AR (Seção 8.2.2.4). Neste caso, os semáforos devem aprender uma política que mapeia os estados (normalmente as filas nas interseções) para ações. Devido ao número de trabalhos que empregam AR no controle semafórico, e às diversas modelagens e técnicas empregadas, sugere-se consultar os *surveys* [Bazzan 2009, Wei et al. 2019, Yau et al. 2017, Noaeen et al. 2022].

Já pelo lado da demanda, entender como um motorista se comporta é fundamental em um sistema de recomendação de rotas e disseminação de informação aos motoristas. Não são muitos os trabalhos que consideram IA neste contexto. Redes neurais são utilizadas em [Dia and Panwai 2014] e em [Barthélemy and Carletti 2017] para prever e guiar,

[18] https://www.processmaker.com/blog/why-ai-is-the-future-of-finance/

[19] https://www.mckinsey.com/industries/financial-services/our-insights/ai-bank-of-the-future-can-banks-meet-the-ai-challenge

respectivamente, a escolha de rota dos motoristas. No caso da pesquisa mais recente, o foco é em: disseminação de informação, comunicação veicular, como aprender a escolher rotas, efeito de mudanças de comportamento da parte dos motoristas na presença de informação, e como disseminar informação de modo a garantir um determinado nível de desempenho do sistema. Para atingir tais objetivos, diversos métodos foram propostos no nosso grupo de pesquisa; para uma visão geral, ver [Bazzan 2022]. Alguns destes trabalhos foram pioneiros ao abordar a disseminação de informação via dispositivos móveis quando o *smartphone* não existia como o conhecemos hoje [Klügl and Bazzan 2004]. Outros métodos envolvem comunicação interveicular [Santos and Bazzan 2021], escolha de rota via AR [Bazzan and Grunitzki 2016], e efeito de recomendação de rotas [Ramos et al. 2018].

Por fim, vale lembrar que também é possível utilizar IA em cenários que combinam controle semafórico com gerenciamento da demanda. De fato, esta integração, tão óbvia quanto importante, tem recebido pouca atenção na literatura. No trabalho de [Wiering 2000] foi um dos pioneiros a tratar motoristas e semáforos aprendendo simultaneamente. Em [Lemos et al. 2018] foi proposta uma abordagem baseada em jogos repetidos (para a classe motorista) e jogos estocásticos (para os semáforos). Por se tratar de naturezas diversas de aprendizado, o artigo também discute os desafios encontrados em termos de AR.

8.8.2. Riscos

De modo geral, os riscos de emprego de IA em aplicações na área de mobilidade urbana são similares a outras já discutidas neste capítulo. Entretanto, entre algumas características específicas, destacam-se as seguintes. Em primeiro lugar, a área de controle semafórico tem a segurança como absoluta prioridade. Desta forma, qualquer método de controle, seja ou não baseado em IA, deve fornecer garantias de que a sinalização não resultará em situações que violem os preceitos fundamentais. Em segundo lugar, no que tange questões de comunicação interveicular, é obviamente fundamental garantir não apenas a privacidade dos envolvidos, mas também a segurança geral do sistema (por exemplo contra ataques maliciosos).

8.8.3. Tendências

Esta seção focou apenas nas questões anteriormente mencionadas – a maioria relacionada a tráfego veicular urbano –. Entretanto, além das questões relacionadas a comunicação interveicular, existem pelo menos três áreas nas quais espera-se avanços significativos pelo uso da IA. A primeira está ligada a otimização do uso da rede de recarga de veículos elétricos. A segunda está, obviamente, relacionada com veículos autônomos e, principalmente, a como acomodar frotas mistas (autônomos e convencionais interagindo no mesmo ambiente). Por fim – e em um horizonte mais concreto de tempo – as aplicações de *mobility as a service* já estão maduras e prontas para serem empregadas em conjuntos de políticas públicas visando dar acesso mais eficiente à populações cada vez mais heterogêneas em suas necessidades de mobilidade.

Parte III: Conclusões e Perspectivas

8.9. Visão geral

Este capítulo apresentou uma introdução aos fundamentos de IA e discutiu aplicações, riscos e tendências em suas múltiplas áreas. Destacamos que tal apresentação não é exaustiva. Há muitos outros conceitos envolvendo IA e muitas outras áreas impactadas que apenas tangenciamos. Este capítulo pode ser visto como um ponto de partida, onde o leitor interessado poderá usar as referências apresentadas para se aprofundar nos tópicos de interesse.

Ao longo do capítulo, é possível ver que a aplicação da IA apresenta riscos em comum nas diferentes áreas. Especificamente, sistemas baseados em aprendizado supervisionado, incluindo aprendizado profundo, possuem questões críticas relacionadas à semântica, explicabilidade, transparência (como e porquê determinada saída foi produzida) e vieses (resultados discriminatórios contra determinados grupos de pessoas). Uma interpretação dos modelos de aprendizado é importante do ponto de vista tecnológico (e de produto) para oferecer garantias sobre o comportamento de um sistema. Ademais, entender exatamente porque um determinado sistema apresenta tal comportamento é um requisito básico de qualquer produto tecnológico. Nesse sentido, modelos de aprendizado profundo, embora tenham apresentados resultados tecnológicos relevantes, não apresentam uma semântica rigorosa (isto é, não são modelos que tenham associados uma interpretação lógico-matemática).

Porém, riscos e acidentes não são exclusividade da IA. Todas as novidades tecnológicas da história da humanidade vieram com seus riscos. Como exemplos: junto com a introdução do automóvel vieram os acidentes automobilísticos e com a eletricidade, vieram os riscos de incêndios causados por curto-circuitos e acidentes por descarga elétrica, entre outros riscos que acompanham tecnologias. Uma questão importante é que nessas tecnologias, um acidente ou evento indesejado ocorre quando "algo vai mal", por falha humana, de hardware, de software, entre outras. Por exemplo, um acidente com automóvel ocorre por falha humana ou em algum de seus componentes; um curto-circuito ocorre por sobrecarga na fiação elétrica. Em contraste, sistemas de IA baseados em aprendizado profundo tem a peculiaridade de que um evento indesejado ocorre mesmo quando "tudo vai bem". Mesmo com toda a implementação correta, e sem falhas no hardware, um sistema como o ChatGPT pode produzir saídas incorretas ou prejudiciais, conforme consta no próprio *disclaimer* em sua página inicial[20].

Um grande tópico de pesquisa envolve, portanto, a identificação e mitigação desses riscos associados à IA. Alguns avanços foram feitos no caminho da explicabilidade [Ribeiro et al. 2016, Lundberg and Lee 2017] e na mitigação de vieses e outros riscos de segurança [Thomas et al. 2019]. Um promissor caminho integrador entre a IA baseada em aprendizado (eficiente, mas pouco transparente e por vezes pouco previsível) e a IA simbólica (menos eficiente até o momento, mas transparente, explicável e previsível) é a abordagem neuro-simbólica, cujos estudos visam, entre outros objetivos, oferecer in-

[20] "Prévia de Pesquisa Gratuita. O ChatGPT pode produzir informações imprecisas sobre pessoas, lugares ou fatos. Versão do ChatGPT de 3 de Agosto", conforme acesso em 22/09/2023.

terpretações (ou explicações, se a posteriori) dos métodos e mecanismos de aprendizado atualmente utilizados em IA, conforme discussão a seguir.

8.10. Integração para lidar com os desafios: A IA Neuro-simbólica

Historicamente, a IA iniciou sua trajetória buscando a integração entre diversas habilidades cognitivas, dentre elas, o raciocínio e o aprendizado. Ambas dimensões são vistas como centrais à ideia de inteligência de máquina, já nos trabalhos originais de Turing, von Neumann, McCulloch, Pitts, entre outros [Turing 1950]. von Neumann, em seus trabalhos iniciais, já identificava a relação entre a lógica intuicionista [von Neumann 1956, d'Avila Garcez et al. 2006] e as redes neurais propostas por [McCulloch and Pitts 1943][21].

A área de IA neuro-simbólica integra os dois principais paradigmas da IA: conexionismo (notadamente associado ao uso de redes neurais artificiais como seu modelo principal) e simbolismo (onde o processo de raciocínio em IA é representado através de lógicas, incluindo diversas modalidades como tempo, espaço, conhecimento e incerteza) [d'Avila Garcez et al. 2007, Lamb et al. 2007, d'Avila Garcez and Lamb 2023]. Tradicionalmente, estas áreas foram desenvolvidas por correntes diversas, por terem fundamentos computacionais e lógicos distintos [Besold et al. 2022, d'Avila Garcez et al. 2009]. A área recebeu certa atenção inicial nos anos 1990 e 2000 [Hinton 1990], quando pesquisadores passaram a desenvolver abordagens neuro-simbólicas que aprendessem a realizar inferência lógica clássica, mesmo que para fragmentos de lógicas de predicados [Audibert et al. 2022, d'Avila Garcez and Zaverucha 1999]. À época, foram desenvolvidos sistemas neuro-simbólicos que aprendiam a computar (fragmentos) de programas escritos em linguagens lógicas, como Prolog. Posteriormente, pesquisadores demonstraram que modelos conexionistas poderiam ser treinados para aprender regras de inferência lógica, notadamente sobre lógicas não-clássicas, permitindo a expressão de multi-modalidades [d'Avila Garcez and Lamb 2003, d'Avila Garcez and Lamb 2006, Lamb et al. 2007], antecipando, de certa forma, a pesquisa atual em grandes modelos de linguagens que visa expressar multimodalidades na interação entre usuários humanos e esses sistemas [Kiros et al. 2014].

As grandes contribuições que a área de IA neuro-simbólica pode oferecer são sumarizadas em artigos recentes, publicados na *Communications of the ACM* [Monroe 2022, Hochreiter 2022]. Monroe ressalta a necessidade de desenvolvimento de uma semântica rigorosa para os modelos de IA, como defendido em [d'Avila Garcez and Lamb 2023, Lamb et al. 2020], enquanto Hochreiter aponta que a forma de desenvolver uma IA ampla, que contemple múltiplas habilidades cognitivas, pode ser melhor atingida através da IA neuro-simbólica, sugerindo a abordagem de redes grafos neurais neuro-simbólicos. Hochreiter cita especificamente o trabalho [Lamb et al. 2020] como sendo promissor para esta linha de pesquisa em IA ampla. É relevante ressaltar que esta necessidade de integração neuro-simbólica foi apontada como promissora em eventos recentes, como nas conferências AAAI e NeurIPS, bem como nos debates organizados pela Montreal AI, de-

[21] [von Neumann 1956, Seção 2] afirma que "It has been pointed out by A. M. Turing [5] in 1937 and by W. S. McCulloch and W. Pitts [2] in 1943 that effectively constructive logics, that is, intuitionistic logics, can be best studied in terms of automata. Thus logical propositions can be represented as electrical networks or (idealized) nervous systems."As referências [5] e [2] na citação são, respectivamente, [Turing 1937] e [McCulloch and Pitts 1943].

nominadas de "AI Debates"[22] números 1, 2 e 3. Nestes eventos, foi apontado que para construir sistemas que representem as duas formas de raciocínio - na forma de "AI Fast and Slow"inspirado em [Kahneman 2011] (ver Seção 8.2) - visando a integração de múltiplas habilidades cognitivas em IA, possíveis abordagens promissoras seriam justamente as adotadas na IA neuro-simbólica. Estas abordagens oferecem a possibilidade do desenvolvimento de modelos com fundamentação rigorosa e transparente do ponto de vista lógico, que se integrados ao aprendizado profundo podem levar a tecnologias de inteligência artificial mais robustas, explicáveis e transparentes, oferecendo maior segurança a todos que fazem uso desta tecnologia de propósito geral [d'Avila Garcez and Lamb 2023].

Mais informações sobre a evolução da IA, em particular IA neuro-simbólica podem ser encontradas em [Audibert et al. 2022, d'Avila Garcez et al. 2009, d'Avila Garcez and Lamb 2023].

8.11. Epílogo

Os impactos sociais e éticos da IA levantam questões preponderantes no debate científico e na grande mídia. Atualmente, inúmeras organizações científicas[23] e academias nacionais de ciência têm debatido o impacto da IA na ciência e na sociedade. Além das entidades científicas, organismos multilaterais, historicamente dedicados a temáticas econômicas e sociais[24], constituíram grupos de trabalho sobre o impacto da IA. Muitas vezes este debate é alimentado por não-especialistas, o que demanda cuidados na interpretação do que é publicado. Mesmo com esta ressalva, é muito relevante perceber que nos últimos 10 a 15 anos, com o crescente impacto da IA, particularmente do aprendizado profundo, os cientistas identificaram uma série de limitações e preocupações com o uso da IA sem curadoria - i.e., sem a própria análise de uso por especialistas.

Por fim, deixamos uma demonstração da capacidade utilitária da IA e como ela pode potencializar a capacidade criativa dos seres humanos. Em resposta ao seguinte *prompt*: "escreva um pequeno texto de conclusão para este capítulo de livro "A Nova Eletricidade": Aplicações, Riscos e Tendências da IA Moderna", a auspiciosa última frase do texto gerado pelo ChatGPT foi: "À medida que fechamos este capítulo, é imperativo olhar para o horizonte da IA com olhos críticos e curiosos, prontos para navegar nas águas emocionantes, porém desafiadoras, deste novo mundo alimentado por algoritmos e dados. O futuro da IA está nas mãos daqueles que a guiam com sabedoria e visão."

Cabe aos seres humanos, portanto, a contínua busca por sabedoria e visão para guiar a IA e todas as tecnologias presentes e futuras para o próprio bem da humanidade.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico. Agradecimentos também

[22] `https://www.quebecartificialintelligence.com/aidebate2/`

[23] e.g. AAAI (Association for the Advancement of Artificial Intelligence), ACM (Association for Computing Machinery), IEEE (Institute of Electrical and Electronic Engineers), e Royal Society, entre outras.

[24] e.g. Fórum Econômico Mundial (WEF), Organização para Cooperação e Desenvolvimento Econômico (OCDE), e as Nações Unidas.

a Cláudio Geyer por comentários no texto e ajuda na revisão.

Referências

[Agrawal et al. 2019] Agrawal, A., Gans, J. S., and Goldfarb, A. (2019). Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2):31–50.

[Al-Zaiti et al. 2023] Al-Zaiti, S. S., Martin-Gill, C., Zègre-Hemsey, J. K., Bouzid, Z., Faramand, Z., Alrawashdeh, M. O., Gregg, R. E., Helman, S., Riek, N. T., Kraevsky-Phillips, K., et al. (2023). Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, pages 1–10.

[Aljameel et al. 2022] Aljameel, S. S., Alomari, D. M., Alismail, S., Khawaher, F., Alkhudhair, A. A., Aljubran, F., and Alzannan, R. M. (2022). An anomaly detection model for oil and gas pipelines using machine learning. *Computation*, 10(8):138.

[Aphirakmethawong et al. 2022] Aphirakmethawong, J., Yang, E., and Mehnen, J. (2022). An overview of artificial intelligence in product design for smart manufacturing. In *2022 27th International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE.

[Audibert et al. 2022] Audibert, R. B., dos Santos, H. L., Avelar, P. H. C., Tavares, A. R., and Lamb, L. C. (2022). On the evolution of A.I. and machine learning: Towards measuring and understanding impact, influence, and leadership at premier A.I. conferences. *arXiv preprint arXiv:2205.13131*.

[Barthélemy and Carletti 2017] Barthélemy, J. and Carletti, T. (2017). A dynamic behavioural traffic assignment model with strategic agents. *Transportation Research Part C: Emerging Technologies*, 85:23–46.

[Baum et al. 2021] Baum, Z. J., Yu, X., Ayala, P. Y., Zhao, Y., Watkins, S. P., and Zhou, Q. (2021). Artificial intelligence in chemistry: current trends and future directions. *Journal of Chemical Information and Modeling*, 61(7):3197–3212.

[Bazzan 2021] Bazzan, A. L. (2021). Contribuições de aprendizado por reforço em escolha de rota e controle semafórico. *Estudos Avançados*, 35(101):95–110.

[Bazzan 2009] Bazzan, A. L. C. (2009). Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multiagent Systems*, 18(3):342–375.

[Bazzan 2022] Bazzan, A. L. C. (2022). Improving urban mobility: using artificial intelligence and new technologies to connect supply and demand. `https://arxiv.org/abs/2204.03570`.

[Bazzan and Grunitzki 2016] Bazzan, A. L. C. and Grunitzki, R. (2016). A multiagent reinforcement learning approach to en-route trip building. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 5288–5295.

[Bazzan and Klügl 2013] Bazzan, A. L. C. and Klügl, F. (2013). *Introduction to Intelligent Systems in Traffic and Transportation*, volume 7 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan and Claypool.

[Bender et al. 2021] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

[Benos et al. 2021] Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., and Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11):3758.

[Berner et al. 2019] Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

[Besold et al. 2022] Besold, T. R., d'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P. M., Hitzler, P., Kühnberger, K., Lamb, L. C., Lima, P. M. V., de Penning, L., Pinkas, G., Poon, H., and Zaverucha, G. (2022). Neural-symbolic learning and reasoning: A survey and interpretation. In Hitzler, P. and Sarker, M. K., editors, *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*, pages 1–51. IOS Press.

[Birlutiu et al. 2017] Birlutiu, A., Burlacu, A., Kadar, M., and Onita, D. (2017). Defect detection in porcelain industry based on deep learning techniques. In *2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 263–270. IEEE.

[Brachman and Levesque 2004] Brachman, R. J. and Levesque, H. J. (2004). *Knowledge Representation and Reasoning*. Elsevier.

[Broda et al. 2004] Broda, K., Gabbay, D., Lamb, L., and Russo, A. (2004). *Compiled Labelled Deductive Systems: A Uniform Presentation of Non-Classical Logics*. Institute of Physics/Research Studies Press, Hertfordshire.

[Brown et al. 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

[Brownstein et al. 2023] Brownstein, J. S., Rader, B., Astley, C. M., and Tian, H. (2023). Advances in artificial intelligence for infectious-disease surveillance. *New England Journal of Medicine*, 388(17):1597–1607.

[Buolamwini and Gebru 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

[Colombelli et al. 2022] Colombelli, F., Kowalski, T. W., and Recamonde-Mendoza, M. (2022). A hybrid ensemble feature selection design for candidate biomarkers discovery from transcriptome profiles. *Knowledge-Based Systems*, 254:109655.

[Dalzochio et al. 2020] Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., and Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, 123:103298.

[d'Avila Garcez and Lamb 2006] d'Avila Garcez, A. and Lamb, L. (2006). A connectionist computational model for epistemic and temporal reasoning. *Neur. Computation*, 18(7):1711–1738.

[d'Avila Garcez et al. 2006] d'Avila Garcez, A., Lamb, L., and Gabbay, D. (2006). Connectionist computations of intuitionistic reasoning. *Theor. Comput. Sci.*, 358(1):34–55.

[d'Avila Garcez and Lamb 2023] d'Avila Garcez, A. and Lamb, L. C. (2023). Neurosymbolic AI: The 3rd Wave. *Artificial Intelligence Review*.

[d'Avila Garcez et al. 2007] d'Avila Garcez, A., Lamb, L. C., and Gabbay, D. M. (2007). Connectionist modal logic: Representing modalities in neural networks. *Theor. Comput. Sci.*, 371(1-2):34–53.

[d'Avila Garcez and Zaverucha 1999] d'Avila Garcez, A. and Zaverucha, G. (1999). The connectionist inductive learning and logic programming system. *Applied Intelligence*, 11(1):59–77.

[d'Avila Garcez and Lamb 2003] d'Avila Garcez, A. S. and Lamb, L. C. (2003). Reasoning about Time and Knowledge in Neural-symbolic Learning Systems. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 921–928. MIT Press.

[d'Avila Garcez et al. 2009] d'Avila Garcez, A. S., Lamb, L. C., and Gabbay, D. M. (2009). *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Springer.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[Dia and Panwai 2014] Dia, H. and Panwai, S. (2014). *Intelligent Transport Systems: Neural Agent (Neugent) Models of Driver Behaviour*. LAP Lambert Academic Publishing.

[dos Anjos et al. 2023] dos Anjos, J. C. S., Matteussi, K. J., Orlandi, F. C., Barbosa, J. L. V., Silva, J. S., Bittencourt, L. F., and Geyer, C. F. R. (2023). A Survey on Collaborative Learning for Intelligent Autonomous Systems. *ACM Comput. Surv.*, 1(1):1–36.

[Dosovitskiy et al. 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Eykholt et al. 2018] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.

[Fagin et al. 1995] Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. (1995). *Reasoning About Knowledge*. MIT Press.

[Fournier-Viger et al. 2021] Fournier-Viger, P., Nawaz, M. S., Song, W., and Gan, W. (2021). Machine learning for intelligent industrial design. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 158–172. Springer.

[Fragapane et al. 2022] Fragapane, G., Ivanov, D., Peron, M., Sgarbossa, F., and Strandhagen, J. O. (2022). Increasing flexibility and productivity in industry 4.0 production networks with autonomous mobile robots and smart intralogistics. *Annals of Operations Research*, 308(1-2):125–143.

[Franklin et al. 2020] Franklin, C. S., Dominguez, E. G., Fryman, J. D., and Lewandowski, M. L. (2020). Collaborative robotics: New era of human–robot cooperation in the workplace. *Journal of Safety Research*, 74:153–160.

[Geffner 2018] Geffner, H. (2018). Model-free, model-based, and general intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI2018*.

[Giri et al. 2019] Giri, C., Jain, S., Zeng, X., and Bruniaux, P. (2019). A detailed review of artificial intelligence applied in the fashion and apparel industry. *IEEE Access*, 7:95376–95396.

[Goodfellow et al. 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[Grisoni et al. 2021] Grisoni, F., Huisman, B. J., Button, A. L., Moret, M., Atz, K., Merk, D., and Schneider, G. (2021). Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Science Advances*, 7(24):eabg3338.

[Gubbi et al. 2022] Gubbi, K. I., Beheshti-Shirazi, S. A., Sheaves, T., Salehi, S., PD, S. M., Rafatirad, S., Sasan, A., and Homayoun, H. (2022). Survey of machine learning for electronic design automation. In *Proceedings of the Great Lakes Symposium on VLSI 2022*, pages 513–518.

[Hamolia and Melnyk 2021] Hamolia, V. and Melnyk, V. (2021). A survey of machine learning methods and applications in electronic design automation. In *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 757–760. IEEE.

[Hart et al. 1968] Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

[Hasan et al. 2021] Hasan, I., Liao, S., Li, J., Akram, S. U., and Shao, L. (2021). Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337.

[Helmert and Domshlak 2009] Helmert, M. and Domshlak, C. (2009). Landmarks, critical paths and abstractions: What's the difference anyway? In *International Conference on Automated Planning and Scheduling*, pages 162–169.

[Hinton 1990] Hinton, G. (1990). Connectionist symbol processing - preface. *Artif. Intell.*, 46(1-2):1–4.

[Hinton et al. 2006] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

[Hochreiter 2022] Hochreiter, S. (2022). Toward a broad AI. *Communications of the ACM*, 65(4):56–57.

[Hochreiter and Schmidhuber 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[Hoffmann 2011] Hoffmann, J. (2011). Everything you always wanted to know about planning: (but were afraid to ask). In *Advances in Artificial Intelligence*, pages 1–13.

[Huang et al. 2019] Huang, P., Lin, C. T., Li, Y., Tammemagi, M. C., Brock, M. V., Atkar-Khattra, S., Xu, Y., Hu, P., Mayo, J. R., Schmidt, H., et al. (2019). Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. *The Lancet Digital Health*, 1(7):e353–e362.

[Ignat et al. 2023] Ignat, O., Jin, Z., Abzaliev, A., Biester, L., Castro, S., Deng, N., Gao, X., Gunal, A., He, J., Kazemi, A., Khalifa, M., Koh, N., Lee, A., Liu, S., Min, D. J., Mori, S., Nwatu, J., Perez-Rosas, V., Shen, S., Wang, Z., Wu, W., and Mihalcea, R. (2023). A PhD student's perspective on research in NLP in the era of very large language models.

[Jiang et al. 2018] Jiang, D., Hao, M., Ding, F., Fu, J., and Li, M. (2018). Mapping the transmission risk of zika virus using machine learning models. *Acta Tropica*, 185:391–399.

[Jumper et al. 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

[Kahneman 2011] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[Kielarova and Pradujphongphet 2023] Kielarova, S. W. and Pradujphongphet, P. (2023). Genetic algorithm for product design optimization: An industrial case study of halo setting for jewelry design. In *International Conference on Swarm Intelligence*, pages 219–228. Springer.

[Kiros et al. 2014] Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal neural language models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–595–II–603. JMLR.org.

[Klügl and Bazzan 2004] Klügl, F. and Bazzan, A. L. C. (2004). Route decision behaviour in a commuting scenario. *Journal of Artificial Societies and Social Simulation*, 7(1).

[Kowalski 1979] Kowalski, R. A. (1979). *Logic for problem solving*. North-Holland.

[Lamb et al. 2007] Lamb, L., Borges, R., and d'Avila Garcez, A. (2007). A connectionist cognitive model for temporal synchronisation and learning. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, page 827–832.

[Lamb et al. 2020] Lamb, L. C., d'Avila Garcez, A. S., Gori, M., Prates, M. O. R., Avelar, P. H. C., and Vardi, M. Y. (2020). Graph neural networks meet neural-symbolic computing: A survey and perspective. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4877–4884. ijcai.org.

[LeCun et al. 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551.

[Lemos et al. 2018] Lemos, L. L., Bazzan, A. L. C., and Pasin, M. (2018). Co-adaptive reinforcement learning in microscopic traffic systems. In *2018 IEEE Congress on Evolutionary Computation, CEC 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–8.

[Li et al. 2017] Li, B.-h., Hou, B.-c., Yu, W.-t., Lu, X.-b., and Yang, C.-w. (2017). Applications of artificial intelligence in intelligent manufacturing: a review. *Frontiers of Information Technology & Electronic Engineering*, 18:86–96.

[Liang et al. 2020] Liang, Y., Lee, S.-H., and Workman, J. E. (2020). Implementation of artificial intelligence in fashion: Are consumers ready? *Clothing and Textiles Research Journal*, 38(1):3–18.

[Liu et al. 2019] Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297.

[Lundberg and Lee 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

[Lynch 2017] Lynch, S. (2017). Andrew Ng: Why AI Is the New Electricity. `https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity`. Acesso em 15/09/2023.

[Mahendran and PM 2022] Mahendran, N. and PM, D. R. V. (2022). A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. *Computers in Biology and Medicine*, 141:105056.

[Marczyk et al. 2023] Marczyk, V. R., Recamonde-Mendoza, M., Maia, A. L., and Goemann, I. M. (2023). Classification of thyroid tumors based on DNA methylation patterns. *Thyroid*, 33(9):1090–1099.

[Martí et al. 2015] Martí, L., Sanchez-Pi, N., Molina, J. M., and Garcia, A. C. B. (2015). Anomaly detection based on sensor data in petroleum industry applications. *Sensors*, 15(2):2774–2797.

[Martin et al. 2019] Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591.

[McCulloch and Pitts 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.

[Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mitchell 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

[Mnih et al. 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

[Monroe 2022] Monroe, D. (2022). Neurosymbolic AI. *Communications of the ACM*, 65(10):11–13.

[Narodytska and Kasiviswanathan 2017] Narodytska, N. and Kasiviswanathan, S. P. (2017). Simple black-box adversarial attacks on deep neural networks. In *CVPR Workshops*, volume 2, page 2.

[Noaeen et al. 2022] Noaeen, M., Naik, A., Goodman, L., Crebo, J., Abrar, T., Abad, Z. S. H., Bazzan, A. L., and Far, B. (2022). Reinforcement learning in urban network traffic signal control: A systematic literature review. *Expert Systems with Applications*, page 116830.

[Obermeyer et al. 2019] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

[Ouyang et al. 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

[Panch et al. 2019] Panch, T., Pearson-Stuttard, J., Greaves, F., and Atun, R. (2019). Artificial intelligence: opportunities and risks for public health. *The Lancet Digital Health*, 1(1):e13–e14.

[Paolanti et al. 2018] Paolanti, M., Romeo, L., Felicetti, A., Mancini, A., Frontoni, E., and Loncarski, J. (2018). Machine learning approach for predictive maintenance in industry 4.0. In *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, pages 1–6. IEEE.

[Papakyriakopoulos et al. 2020] Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., and Marco, F. (2020). Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457. Association for Computing Machinery.

[Phakhounthong et al. 2018] Phakhounthong, K., Chaovalit, P., Jittamala, P., Blacksell, S. D., Carter, M. J., Turner, P., Chheng, K., Sona, S., Kumar, V., Day, N. P., et al. (2018). Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: application of classification tree analysis. *BMC Pediatrics*, 18:1–9.

[Pivetta et al. 2023] Pivetta, M. V. L., Simon, A. H., Costa, M. M., Abel, M., and Carbonera, J. L. (2023). A systematic evaluation of machine learning approaches for petroleum production forecasting. In *IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 768–774. IEEE.

[Prates et al. 2020] Prates, M. O., Avelar, P. H., and Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

[Radford et al. 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR. ISSN: 2640-3498.

[Rahmanifard and Plaksina 2019] Rahmanifard, H. and Plaksina, T. (2019). Application of artificial intelligence techniques in the petroleum industry: a review. *Artificial Intelligence Review*, 52(4):2295–2318.

[Rajpurkar et al. 2022] Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1):31–38.

[Ramos et al. 2018] Ramos, G. de. O., Bazzan, A. L. C., and da Silva, B. C. (2018). Analysing the impact of travel information for minimising the regret of route choice. *Transportation Research Part C: Emerging Technologies*, 88:257–271.

[Ribeiro et al. 2021] Ribeiro, J., Lima, R., Eckhardt, T., and Paiva, S. (2021). Robotic process automation and artificial intelligence in industry 4.0–a literature review. *Procedia Computer Science*, 181:51–58.

[Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

[Roth et al. 2022] Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. (2022). Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328.

[Rumelhart et al. 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[Russell et al. 2015] Russell, S., Hauert, S., Altman, R., and Veloso, M. (2015). Ethics of artificial intelligence: Four leading researchers share their concerns and solutions for reducing societal risks from intelligent machines. *Nature*, 521:415–418.

[Russell and Norvig 2020] Russell, S. and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson.

[Santos and Bazzan 2021] Santos, G. D. dos. and Bazzan, A. L. C. (2021). Sharing diverse information gets driver agents to learn faster: an application in en route trip building. *PeerJ Computer Science*, 7:e428.

[Schuster and Paliwal 1997] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

[Schwalbe and Wahl 2020] Schwalbe, N. and Wahl, B. (2020). Artificial intelligence and the future of global health. *The Lancet*, 395(10236):1579–1586.

[Serradilla et al. 2022] Serradilla, O., Zugasti, E., Rodriguez, J., and Zurutuza, U. (2022). Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. *Applied Intelligence*, 52(10):10934–10964.

[Sharir et al. 2020] Sharir, O., Peleg, B., and Shoham, Y. (2020). The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*.

[Silver et al. 2017a] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017a). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

[Silver et al. 2017b] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017b). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.

[Stojanovic et al. 2016] Stojanovic, L., Dinic, M., Stojanovic, N., and Stojadinovic, A. (2016). Big-data-driven anomaly detection in industry (4.0): An approach and a case study. In *2016 IEEE International Conference on Big Data*, pages 1647–1652. IEEE.

[Sutton and Barto 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. The MIT Press, second edition.

[Thomas et al. 2019] Thomas, P. S., Castro da Silva, B., Barto, A. G., Giguere, S., Brun, Y., and Brunskill, E. (2019). Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004.

[Toorajipour et al. 2021] Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., and Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, 122:502–517.

[Turing 1937] Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265.

[Turing 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

[Van der Schaar et al. 2021] Van der Schaar, M., Alaa, A. M., Floto, A., Gimson, A., Scholtes, S., Wood, A., McKinney, E., Jarrett, D., Lio, P., and Ercole, A. (2021). How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning*, 110:1–14.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[Vinyals et al. 2019] Vinyals, O., Babuschkin, I., Czarnecki, W., Mathieu, M., Dudzik, A., Chung, J., Choi, D., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J., Jaderberg, M., and Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350—-354.

[von Neumann 1956] von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata Studies*, 34:43–98.

[Wang and Luo 2019] Wang, L. and Luo, M. (2019). Machine learning applications and opportunities in ic design flow. In *2019 International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*, pages 1–3. IEEE.

[Warren et al. 2023] Warren, D. S., Dahl, V., Eiter, T., Hermenegildo, M. V., Kowalski, R. A., and Rossi, F., editors (2023). *Prolog: The Next 50 Years*, volume 13900 of *Lecture Notes in Computer Science*. Springer.

[Watkins 1989] Watkins, C. (1989). *Learning from Delayed Rewards*. PhD thesis, University of Cambridge.

[Watkins and Dayan 1992] Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.

[Wei et al. 2019] Wei, H., Li, Z., Xu, N., Zhang, H., Zheng, G., Zang, X., Chen, C., Zhang, W., Zhu, Y., and Xu, K. (2019). Colight: Learning network-level cooperation for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1913–1922. Association for Computing Machinery.

[Wiering 2000] Wiering, M. (2000). Multi-agent reinforcement learning for traffic light control. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, pages 1151–1158.

[Wong et al. 2021] Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070.

[World Health Organization 2021] World Health Organization (2021). Ethics and governance of artificial intelligence for health: WHO guidance.

[Xu et al. 2019] Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., Beaty, K. A., Dehan, E., and Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Human Genetics*, 138(2):109–124.

[Yau et al. 2017] Yau, K.-L. A., Qadir, J., Khoo, H. L., Ling, M. H., and Komisarczuk, P. (2017). A survey on reinforcement learning models and algorithms for traffic signal control. *ACM Comput. Surv.*, 50(3).

[Yu et al. 2018] Yu, K.-H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731.

[Zhang et al. 2019] Zhang, X., Zhou, T., Zhang, L., Fung, K. Y., and Ng, K. M. (2019). Food product design: a hybrid machine learning and mechanistic modeling approach. *Industrial & Engineering Chemistry Research*, 58(36):16743–16752.

[Zhu et al. 2021] Zhu, X., Ninh, A., Zhao, H., and Liu, Z. (2021). Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. *Production and Operations Management*, 30(9):3231–3252.

[Zipfel et al. 2023] Zipfel, J., Verworner, F., Fischer, M., Wieland, U., Kraus, M., and Zschech, P. (2023). Anomaly detection for industrial quality assurance: A comparative evaluation of unsupervised deep learning models. *Computers & Industrial Engineering*, 177:109045.