IA Generativa: Fundamentos, Usos e Riscos

Rogério de Oliveira

rogerio.oliveira@{maua.br | mackenzie.br}

Resumo

Este capítulo explora diversos aspectos da Inteligência Artificial (IA), com foco em IA generativa, suas aplicações, riscos e benefícios no cenário atual. Procuro transcrever aqui, o mais próximo possível, a palestra de mesmo título que compôs a segunda parte da 3ª sessão de palestras da Cátedra de Design com o tema Paisagens Híbridas, em 13 de março de 2024. Apenas que, por limitações para a versão impressa, há um bastante menor número figuras, embora a palestra e seus slides possam ser acessados de modo integral em Oliveira (2024). Tendo como público especialistas na área de design e interessados em geral, apresento na primeira parte, após a introdução, alguns dos principais fundamentos de inteligência artificial que acredito serem úteis para entender tanto as capacidades como as limitações dessas tecnologias que inclui um breve histórico da IA, conceitos de aprendizado supervisionado, modelos neurais e aprendizado profundo, até os recentes modelos de IA generativa e grandes modelos de linguagem. Em seguida, com base nesses fundamentos, discuto alguns dos desafios tecnológicos e sociais da IA, em particular dos modelos generativos, para concluir com seus usos e capacidades. Ao final, enfatizo a necessidade de colaborações intersetoriais para mitigação dos riscos e para garantirmos que a IA traga o benefício que promete à sociedade.

1. Introdução

A inteligência artificial (IA) vem transformando diversos setores da sociedade, desde a medicina e áreas de saúde, até as finanças, passando pela agricultura, o comércio eletrônico, produção cultural e o design, com impactos visíveis também no dia a dia das pessoas. Com o avanço das técnicas de aprendizado profundo (*deep learning*) e a criação de modelos neurais complexos, a IA generativa emergiu como uma das áreas mais inovadoras e promissoras. Esta palestra (Oliveira, 2024), aqui transcrita e adaptada, visa

discutir os fundamentos da IA generativa, suas capacidades e usos, bem como os desafios e riscos recentes associados à sua implementação.

2. Fundamentos de Inteligência Artificial

Acredito que os conceitos que apresento a seguir, apesar de às vezes técnicos, estão ao alcance do público em geral e são fundamentais para que, mesmo os não especialistas, possa entender a capacidades e limitações das tecnologias de IA que são hoje empregadas. Questões como “por que os modelos erram?”, “como adquirem conhecimento?”, “que fatores contribuíram para chegarmos até aqui?” e outras questões fundamentais, podem ser melhor compreendidas através desses fundamentos e permitirão que o não especialista interessado possa refletir essas questões por si próprio, mais independente de fontes e opiniões secundárias, muitas vezes viesadas pela paixão ou pré-conceitos que as mudanças disruptivas sempre carregam.

2.1 História e Evolução

Modelos de redes neurais, base da IA moderna, não são novidade. Desde os anos 1950, redes neurais são conhecidas e estudadas. Mas foi somente nos últimos dez ou quinze anos, que observamos uma revolução disruptiva, com o surgimento das redes neurais profundas, de modelos de grande escala e sua popularização com amplo acesso pelos indivíduos. A principal diferença para essa mudança disruptiva com relação aos modelos anteriores, reside na forma de representação dos dados e das informações, mais complexa, mas também em uma grande disponibilidade de dados e capacidades de processamento, que permitiram capacidades inéditas dos novos modelos.

É importante destacar que a IA é um amplo campo de estudo que envolve diversas técnicas que buscam reproduzir o comportamento inteligente do ser humano. Ela envolve desde a robótica e algoritmos de busca para solução de problemas complexos, como problemas de determinação de rotas e menores caminhos, até técnicas de algoritmos genéticos (muito aplicados em otimização),

e sistemas de produção baseados em regras e lógica de primeira ordem para produzir conhecimento (o clássico, das regras: Todo homem é mortal; Sócrates é homem; deriva-se: Sócrates é mortal). Mais recentemente, entretanto, o termo IA vem sendo mais aplicado para se referir a um tipo específico de técnicas de IA e, é muito provável, que se você encontrar um artigo, notícia ou publicidade relacionado à IA, que de fato a notícia se refira a uma aplicação de aprendizado de máquina ou, mais provavelmente ainda, a um modelo de IA baseado em aprendizado profundo (redes neurais profundas, ou *deep learning*). A figura 1 mostra esse relacionamento de inclusão dentre os diferentes conceitos.

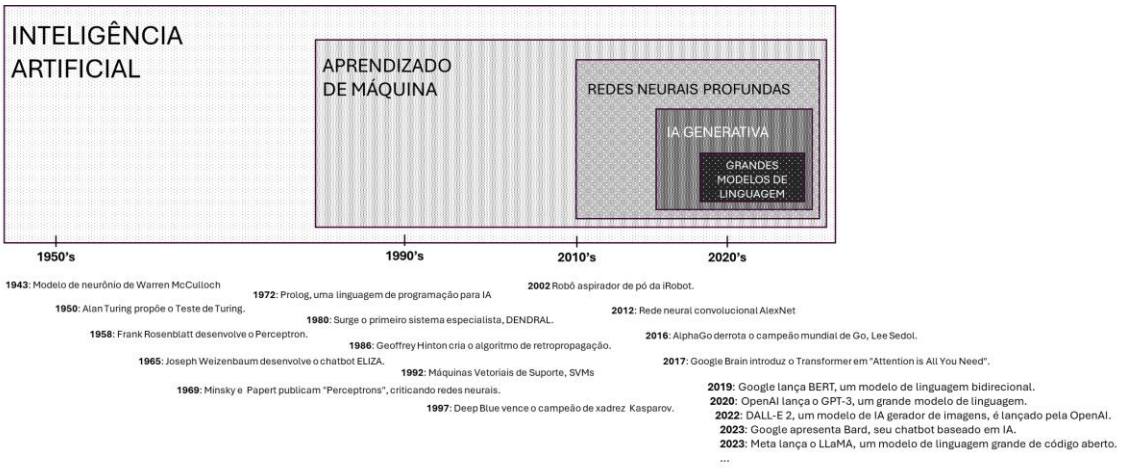

Figura 1. A inteligência artificial é um campo amplo que inclui grande número de técnicas e conceitos que visam reproduzir o comportamento inteligente do ser humano. O aprendizado de máquina é apenas uma dessas técnicas e que vem tendo bastante sucesso nos anos recentes, auxiliado pela gigantesca disponibilidade de dados, capacidade de processamento e novos modelos de redes neurais.

2.2 Aprendizado Supervisionado

O aprendizado de máquina é um campo da inteligência artificial em que os sistemas aprendem e melhoram seu aprendizado com base em dados (experiências anteriores) sem serem explicitamente programados (Mitchell, 1997). Não à toa, esses sistemas são assim bastante beneficiados com a grande disponibilidade de dados produzida pela transformação digital e a internet. Existem ainda diferentes paradigmas de aprendizado de máquina, incluindo o aprendizado supervisionado, não supervisionado e por reforço.

Por hora, o que paradigma que mais nos interessa é o do aprendizado supervisionado, pois ele permite entender como são “treinados” os recentes grandes modelos de linguagem, como o ChatGPT-3/4 (Open AI), Bert/Bart/Gemini (Google), Llama (Meta) e outros modelos generativos de texto ou mesmo imagens e vídeos.

O aprendizado supervisionado é uma técnica de aprendizado de máquina em que um algoritmo é treinado usando dados de entrada *rotulados*. Você pode entender esse conjunto de dados como um conjunto de exemplos de entradas e saídas (os rótulos), que podem ser classes de dados, valores ou, como nos modelos de linguagem a próxima palavra em uma sequência de palavras em um texto. O objetivo do algoritmo é aprender uma função que melhor mapeia as entradas às saídas desejadas com base nos dados fornecidos. Durante o treinamento, o algoritmo ajusta seus *parâmetros* para minimizar a diferença entre suas previsões e as saídas verdadeiras no conjunto de dados de treinamento. Esse processo permite que o modelo “aprenda” a fazer previsões de novos dados, que não fazem parte do conjunto de treinamento e que, portanto, não nunca foram “vistos” antes. Ver figura 2.

2.2 Redes Neurais

Existem diferentes modelos, ou algoritmos, de aprendizado de máquina supervisionado, como árvores de decisão, máquinas de vetores de suporte, k-vizinhos mais próximos, regressão logística etc. Esses algoritmos são amplamente conhecidos e podem ser encontrados em diversos livros texto como em Bishop e Nasrabadi (2006) e Mitchell (1997), além de haver inúmeras bibliotecas/pacotes abertos disponíveis com implementações e que popularizaram bastante o uso desses algoritmos, como por exemplo a biblioteca scikit-learn (2024).

Dentre todos esses modelos, entretanto, os modelos de redes neurais vêm se destacando desde o início da década de 2000 pela sua capacidade de previsão e solução de problemas com dados altamente complexos e de grandes dimensões, como dados de texto, imagem, áudio e vídeo, que surgem em

problemas de direção e navegação autônoma, diagnóstico em exames médicos por imagem, transcrição e tradução de linguagem natural (texto ou áudio), e atingindo resultados sem precedentes comparados aos outros modelos.

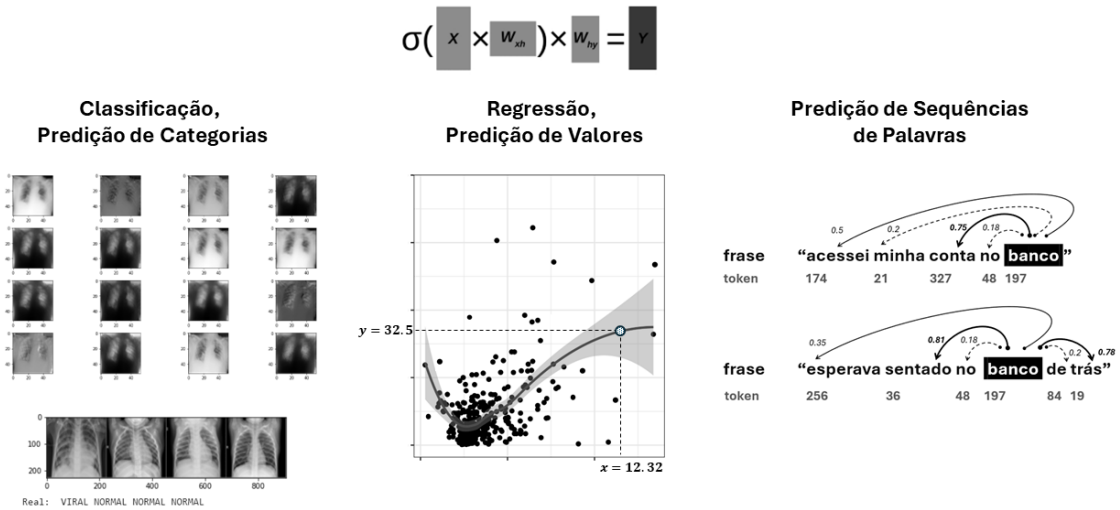

Figura 2. Diferentes tarefas de previsão que podem ser realizadas por modelos de aprendizado supervisionado. A predição de classes ou categorias (valores discretos), por exemplo, identifica a partir de um conjunto inicial de imagens de radiografias de pulmão rotuladas como VIRAL ou NORMAL, se uma nova radiografia apresenta um pulmão saudável (NORMAL) ou comprometido (VIRAL); A predição de valores, valores contínuos de uma “função”, $f: \boldsymbol{x} \rightarrow y$, como preços de imóveis (y) a partir de suas características como número de dormitórios, metros quadrados etc. (representados por $\boldsymbol{x}$), ou a temperatura ou nível de emissões de poluentes, a partir de dados históricos; E a predição da próxima palavra em uma sequência de palavras de um texto, como são treinados os atuais grandes modelos de linguagem.

Dentre todos esses modelos, entretanto, os modelos de redes neurais vêm se destacando desde o início da década de 2000 pela sua capacidade de previsão e solução de problemas com dados altamente complexos e de grandes dimensões, como dados de texto, imagem, áudio e vídeo, que surgem em problemas de direção e navegação autônoma, diagnóstico em exames médicos por imagem, transcrição e tradução de linguagem natural (texto ou áudio), e atingindo resultados sem precedentes comparados aos outros modelos.

As redes neurais artificiais são compostas por neurônios “artificiais” que nada mais são do que artefatos de software (você pode entender como pequenos programas) que recebem entradas, processam e produzem saídas (ver figura 3).

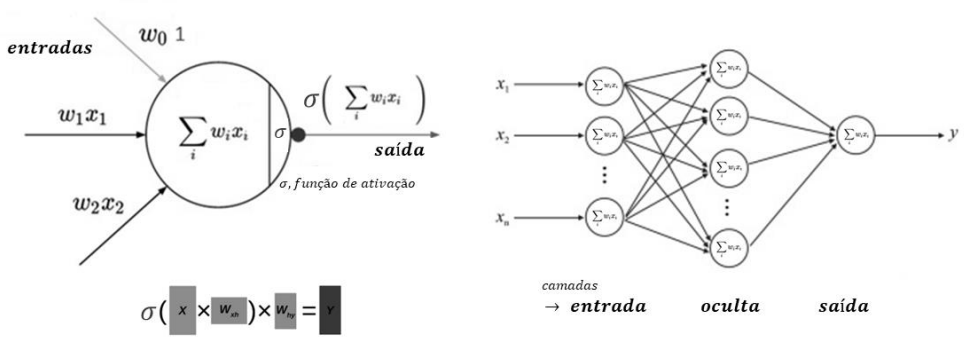

Figura 3. À esquerda, esquema do funcionamento de um neurônio artificial (perceptron) como empregado em redes neurais, incluindo redes neurais muito complexas e com centenas de milhares de elementos, como as empregadas pelos atuais grandes modelos de linguagem (ChatGPT-4, Llama, Gemini). O processamento consiste em aplicar uma função de ativação, como a função logística, ReLu ou Softmax, a uma combinação linear das entradas, para produzir uma saída. O treinamento, ou o aprendizado, consiste em ajustar, através de várias iterações, os pesos do neurônio (w_i) para aproximar ao máximo suas saídas das saídas desejadas. Em geral, isso é implementado por um artefato de software, como parte um programa de software maior. À direita, a combinação de vários elementos perceptron em camadas, que podem chegar aos bilhões em grandes modelos, fornece a esses sistemas a capacidade de solução de problemas complexos como a direção autônoma, classificação e geração de imagens, e a transcrição e tradução de textos.

O ajuste dos "pesos" desses neurônios ou seus parâmetros, através de um processo de treinamento, permite à rede aprender a executar diversas tarefas, desde classificação de imagens até o processamento de linguagem natural. Por exemplo, a função AND, $f:(0,0) \rightarrow 0; f:(0,1) \rightarrow 0;\ f:(1,0) \rightarrow 0;\ f:(1,1) \rightarrow 1$ que somente leva entradas de mesmo valor para o valor 1 (TRUE), pode ser aprendida por um neurônio perceptron, com os pesos ajustados para $w_0 = -0.3, w_1 = 0.2$ e $w_2 = 0.2$. Para chegarmos a esses valores inicia-se o neurônio com valores de pesos arbitrários, calcula-se a saída e, dependendo do "erro" obtido (a diferença entre a saída obtida e a desejada) ajustam-se os valores dos pesos. O processo é repetido até que o valor da saída fique próximo à saída desejada. Esse processo de ajuste é conhecido como algoritmo de retropropagação, ou *backpropagation*, e é um processo bastante complexo e que envolve grande número de operações de cálculo. A ideia é a mesma quando no lugar de um único neurônio temos milhares desses elementos conectados (ver figura 3), embora o ajuste (aumento ou a diminuição dos pesos) envolva uma série de cálculos ainda mais complexos para que o ajuste seja o mais eficiente possível, com o treinamento convergindo para o valor desejado

mais rapidamente. Este processo de ajuste é fundamental para o funcionamento de modelos como o ChatGPT-4, Llama ou Gemini, que utilizam bilhões de parâmetros treináveis, e isso permite entender por que o aprendizado desses modelos requer altas capacidades de processamento.

Para uma experiência com o aprendizado de modelos de redes neurais, mas sem a necessidade programar ou mesmo ter conhecimentos de programação, você pode acessar https://playground.tensorflow.org/ (2024) e experimentar o ajuste de pesos de um modelo neural para problemas simples de até duas entradas e uma saída de dados. Para saber mais e em profundidade esses modelos neurais Goodfellow et. al (2016) e Zhang et al. (2023) são ótimas referências e estão disponíveis online.

2.3 Modelos de Representação dos Dados

O processamento dos elementos de uma rede neural é então basicamente um cálculo de combinações de valores de entradas a que se aplica uma função (chamada função de ativação), e que representamos por $\sigma(x \times w_{xh}) \times w_{hy} = y$. Sendo um cálculo efetuado sobre as entradas é necessário que as entradas sejam valores numéricos para que esses cálculos possam ser realizados. Assim, imagens, textos, ou dados de um imóvel que queremos introduzir em um modelo de rede neural precisam de alguma forma estarem todos representados por dados numéricos. Isso pode parecer imediato para dados como os metros quadrados ou o número de dormitórios de um imóvel, ou ainda, para quem tem familiaridade com a representação de *pixels* de imagens em computadores (valores de 0-255 que são combinados na representação RGB de cores). Mas bastante menos óbvia quando falamos da linguagem natural (textos, falas ou mesmo a descrição textual das características de um imóvel). Embora saibamos que um texto em linguagem natural pode ter todas as suas letras representadas por caracteres armazenados como números em um computador (por exemplo, a letra A é armazenada com o valor 65 em um byte em sua representação ASCII, e padrão da maioria dos computadores atuais), esse tipo de representação não permite representar quaisquer significados das palavras ou frases, diferenciar os diferentes usos de uma

palavra como ‘manga’ (a fruta, a manga da camisa, ou o nome dado a desenhos de origem japonesa), ou representar as complexas relações de cada termo em um texto (sujeito, advérbio etc.).

Construir uma representação que capture aspectos como esses da linguagem natural é bastante mais complexo do que simplesmente fornecer uma representação para armazenamento (como o padrão ASCII). Mas, para os propósitos dessa introdução é suficiente entendermos como uma representação pode ser útil para a “compreensão” da linguagem, sem termos a pretensão de construí-la aqui. Nos atuais grandes modelos de linguagem, textos são representados por sequencias de unidades chamadas “tokens”. A tokenização pode empregar letras, palavras ou parte das palavras sendo esta última forma o método mais comum. Assim, o GPT ou Bert (dois esquemas de tokenização bastante empregados) pode, por exemplo, dividir a frase "*Esperava sentado no banco de trás*" no tokens ["Esper", "ava", "sent", "ado", "no", "banc", “o”, “de”, "trás”], associando a cada token um código numérico. Essa representação é útil para construir um modelo em que a totalidade dos pesos da rede neural irá mapear a relação entre os diferentes termos (tokens) de um texto. Esse mapeamento pode, então, ser empregado para determinar o termo (token) mais provável em uma sequência de termos (tokens). Representamos isso de forma bastante simplificada na figura 2, na Predição de sequências de palavras. Os números da frase representam os tokens (das palavras para simplificar) e os números acima, com setas, os “pesos” (um único peso “final”, embora o modelo contenha milhares de pesos para o mapeamento da relação entre os diferentes tokens) que envolvem o quanto os diferentes termos estão relacionados. Assim, ao buscar completar a frase “acessei minha conta no ___”, a palavra “banco” tem um peso (uma probabilidade maior) que outras palavras, como “computador” ou “aplicativo” (outras possibilidades talvez com probabilidade menor). No exemplo seguinte (logo abaixo, na figura 2), representamos como o modelo ainda pode mapear relações bidirecionais que encontramos em um texto. Essa vizinhança de termos em torno de uma palavra, é chamada de n-grama, onde n é o número de termos (tokens) vizinhos à direita e à esquerda de um termo que são considerados.

Em resumo, a tokenização transforma os textos em vetores numéricos e, com essa representação, os modelos buscam capturar os contextos semânticos e as relações complexas das palavras. Para maiores detalhes sobre o processo de tokenização o leitor pode consultar Mielke et al. (2021). Para imagens, as redes convolucionais são modelos que mapeiam a vizinhança de cada pixel, e desempenham um papel semelhante ao da representação de textos que vimos aqui.

2.4 Aprendizado Profundo e Modelos de Atenção

O modelo de neurônio artificial empregado nas redes neurais profundas difere muito pouco do perceptron que já era empregado nas redes neurais das décadas de 80 e 90. Embora o aprendizado profundo esteja fortemente associado a redes neurais com muitas camadas, a principal diferença entre os primeiros modelos neurais e os modelos de *deep learning* atuais é a capacidade desses últimos modelos criarem representações internas ('profundas') e hierárquicas para dados muito complexos, como textos e imagens (Goodfellow et al., 2016).

Vale ainda citarmos um último avanço na arquitetura desses modelos que consiste no surgimento da arquitetura Transformers (Waswani et al., 2017). Até o momento essa é hoje a arquitetura dominante de todos grandes modelos de linguagem (incluídos os empregados em imagens), assim como em ferramentas de busca, de tradução e de transcrição de textos. Os Transformers (Waswani et al., 2017), introduzido pela Google em 2017, revolucionou os sistemas de processamento de linguagem natural ao substituir os modelos de redes neurais recorrentes antes empregados, por uma arquitetura baseada no mecanismo conhecido como “mecanismo de atenção”. Intuitivamente, esse mecanismo de atenção permite que o modelo receba todas as palavras de uma única vez e identifique os termos importantes relacionados no texto sem a necessidade de um processamento sequencial das palavras. Algo como olhar para esse parágrafo (sem lê-lo do início ao fim) e identificarmos as palavras Transformers e Google como os mais relevantes, sem lermos cada uma das palavras. De fato, nosso cérebro já “aprendeu” e tem pesos diferentes para diferentes palavras

atribuindo, por exemplo, pesos menores a artigos, preposições que para substantivos e nomes. Se você era usuário de ferramentas de transcrição e tradução de textos na internet na última década, deve ter notado a grande avanço que essa tecnologia proporcionou. Esse mecanismo de atenção também permitiu o processamento das entradas de forma paralela, oferecendo assim uma escalabilidade desses modelos sem precedentes.

2.5 Lição desses Fundamentos

Não é esperado, na introdução desses conceitos, que o leitor não especialista venha a entender completamente os complexos mecanismos por traz dessas tecnologias. Mas o entender alguns desses fundamentos é essencial para compreender as limitações, capacidades e riscos dessa tecnologia e compreender melhor as transformações que estamos acompanhando.

Em síntese, você pode apreender do que vimos até aqui que:

a. O avanço recente dessas tecnologias se deve basicamente a 3 fatores: i) uma grande abundância de dados disponível, gerada pela recente transformação digital e internet; ii) uma abundância ampla de recursos computacionais, requeridos para o treinamento de grandes modelos; iii) o avanço de novas arquiteturas de aprendizado profundo (redes neurais) nos últimos anos.

b. Os modelos de aprendizado profundo são modelos probabilísticos, que criam representações das relações bastante complexas dos dados, e são úteis para prever novos dados com base na distribuição (de probabilidades) dos dados que foram empregados para treinar os modelos.

Entender claramente esses princípios básicos ajuda-nos a melhor responder perguntas como: porque a IA generativa pressiona o consumo global de energia; porque ela reforça concepções e preconceitos já existentes; ou, se os modelos atuais de fato “pensam” como o ser humano. Alguns dos termos e

modelos que vimos aqui, e como eles se relacionam são apresentados na figura 4.

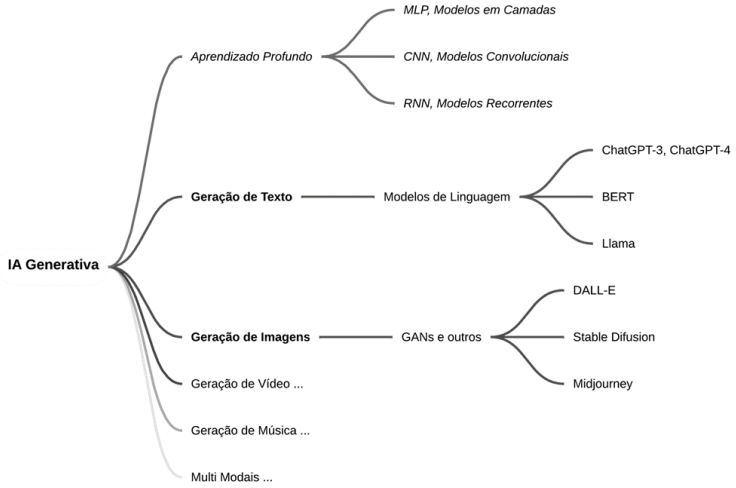

Figura 4. Diagrama de como alguns dos principais conceitos e modelos da IA Generativa estão relacionados.

3. Usos e Capacidades da IA Generativa

Com base no que aprendemos sobre os modelos de IA até aqui, podemos discutir alguns de seus principais usos e capacidades.

3.1 O argumento do quarto chinês

Imagine uma pessoa que não fala chinês sentada em um quarto fechado onde há um conjunto de regras (um grande livro ou manual) que permite a ela correlacionar quaisquer sequências de símbolos em chinês de entrada com sequências de símbolos em chinês de saída. Para uma outra pessoa que entregue frases em chinês por debaixo da porta do quarto e receba em resposta frases corretas, todas em chinês, para cada frase, irá parecer que a pessoa dentro do quarto compreende completamente o chinês.

Esse é o argumento do quarto chinês de Searle (1980) que nos desafia a refletir sobre a natureza da compreensão e da consciência dos sistemas de IA. Grandes modelos de linguagem vêm demonstrando capacidades

impressionantes ao responderem questões complexas de forma bastante convincente (embora nem sempre correta), mas a verdadeira compreensão pode estar bastante além de simplesmente manipular símbolos. Como vimos, assim como no quarto chinês, os grandes modelos de linguagem manipulam símbolos (palavras, frases, tokens) com base em pesos “aprendidos” a partir de um grande volume de dados (como as páginas da internet), mas não têm uma compreensão real de seu significado subjacente. Eles produzem respostas que parecem convincentes e compreensivas aos humanos, baseadas na distribuição de probabilidades dos dados, mas não possuem consciência ou entendimento desses conteúdos (embora haja uma certa discussão sobre isso). Se um modelo fosse treinado com diversos textos que afirmassem que a “terra é plana”, e se fosse perguntado a ele “a terra é redonda?”, o modelo poderia responder que não, o que explica muitos dos problemas (mas não todos) que encontramos nas respostas desses modelos.

Essa limitação, dada pela forma de como os modelos de IA generativa são construídos, é talvez a mais importante de compreendermos para entender a capacidade desses modelos e sua limitação em tratar alguns problemas, como problemas de lógica e raciocínio, em que muitas vezes esses modelos falham.

3.2 Aplicações atuais

Não obstante isso, a IA generativa, tem um amplo espectro prático de aplicações como mostra a Tabela 1, que envolve desde áreas de criação de conteúdo, design e entretenimento, até áreas de desenvolvimento de aplicações, jogos e robótica, com uma grande versatilidade.

Dentre essas aplicações destacamos de Huang e Grady (2022):

a. *Copywriting.* Com a produção de conteúdo personalizado para web e e-mails, essencial para estratégias de vendas, marketing e suporte ao cliente.

Tabela 1 - Quadro de algumas das aplicações mais comuns dos grandes modelos de linguagem. (Adaptado de Huang e Grady, 2022).

Camada de Aplicação	Marketing (conteúdo) Vendas (emails) Suporte (chat | emails) Escrita em Geral Produção de notas Outros	Geração de Código Documentação Texto para SQL Criação Automática de Aplicações Web	Geração de Imagem Consumo e Mídia Social Mídia Publicitária Design	Síntese de Voz	Edição e Geração de Vídeo	Modelos e Cenas 3D	Jogos RPA (Robótica) Música Áudio Biologia e Química
	TEXTO	**CÓDIGO**	**IMAGEM**	**VOZ**	**VÍDEO**	**3D**	**Outros**
Modelos	Open AI GPT-3 | 4 Google Gemini	Open AI GPT-3 | 4 MS Copilot	Open Ai Dall E-2 Stable Diffusion	Open AI	Meta Make-A-Video	DreamFusion Midjourney	-

b. ***Assistentes de Escrita Específicos para Verticais.*** Diversos assistentes de escrita específicos para diferentes mercados como por exemplo, contratos legais, roteiros, descritivo de produtos, licitações etc.

c. ***Geração de Código.*** Ferramentas como o GitHub e o Microsoft Copilot já geram hoje cerca de 40% do código onde essas ferramentas são empregadas.

d. ***Geração de Arte.*** Exploração de temas e estilos, que podem ser empregados em diferentes mercados como estampas de tecidos, material publicitário etc.

e. ***Jogos e filmes.*** O uso da linguagem natural para criar cenas ou modelos complexos é ainda distante, mas opções vêm sendo criadas também nesta área.

f. ***Mídia/Publicidade.*** Automatizar do trabalho de agências e otimizar mensagens publicitárias em tempo real para consumidores.

g. ***Design.*** Prototipagem de produtos desde utensílios e móveis, até casas e prédios, a partir de um processo iterativo e facilitado por renderizações de alta fidelidade e esboços a partir de prompts.

Adicionalmente, acrescento que esses modelos também vêm sendo utilizados para extrair e analisar grandes volumes de dados, identificando padrões e sendo particularmente útil em setores de pesquisa científica, finanças e marketing, em que a análise de dados desempenha um papel crucial.

3.3 Aplicações futuras

Segundo o Gartner (2022), um grande instituto de pesquisas de mercado de tecnologia e inovação, o uso de IA e dos modelos generativos encontra-se hoje em um momento em que as expectativas com relação a esses modelos estão bastante infladas devendo haver um decréscimo das expectativas para os próximos anos (ver, o Hype Cycle de inovação para a IA, em Gartner, (2022)).

Não obstante esse entusiasmo, parece que ainda estamos nas aplicações talvez mais óbvias, ou imediatas, desses modelos. O Gartner (2023), por exemplo, projeta casos de uso e aplicações talvez bastante mais impactantes desses modelos para os próximos anos, trazendo inovações para setores menos óbvios, como a indústria farmacêutica e a ciência de materiais (ver Tabela 2).

Dentre as previsões, estão que até 2025, 30% das mensagens de marketing deverão estar sendo produzidas sinteticamente e que até 2030 haverá filmes de grande sucesso (“blockbusters”) sendo lançados com 90% do seu conteúdo produzido por IA generativa, sendo evidente o crescimento exponencial que vemos de criação e uso desses modelos.

3. Desafios e Riscos da IA

Claramente, por sua abrangência nos mais diferentes setores, sua velocidade de adesão e uso, e as mudanças que trazem, os impactos desses modelos, seja no mercado, seja no dia a dia das pessoas ou na sociedade, deverá ser muito grande, e traz muitos desafios e riscos associados. Em que pese ainda, que estamos falando de uma tecnologia de fronteira em que ainda há muitas dúvidas e muitas mudanças ainda podem ocorrer. Mas, além de enumerar e dar exemplos de casos de riscos, parece útil estruturarmos os principais riscos dessa tecnologia.

Tabela 2 – Casos de Uso e Aplicações futuras da IA generativa segundo o Gartner (2023). (Adaptado de Gartner, 2023)

	Segmentos da Indústria							
Casos de Uso	Manufatura de Automóveis e outros Veículos	Mídia	Arquitetura e Engenharia	Energia e *Utilities*	Provedores de Saúde	Manufatura de Produtos Eletrônicos	Manufatura em Geral	Indústria Farmacêutica
Desenho | Projeto de Novas Drogas								X
Ciência dos Materiais	X			X		X		
Desenho de Circuitos (Chips)						X		
Produção de Dados Sintéticos	X		X	X	X	X	X	X
Design Generativo (parcial)	X		X				X	

O Departamento para Ciência, Inovação e Tecnologia do Reino Unido, encomendou o estudo *Frontier AI: capabilities and risks – discussion paper A discussion paper on the capabilities of, and risks from, frontier AI* (Gov.uk, 2023). Ele organiza os riscos dessas novas tecnologias, e parece ser bastante útil para os nossos propósitos. Empregaremos em toda a discussão que segue, a estrutura empregada por esse trabalho, para melhor compreendermos os riscos associados à IA e para aproveitar com segurança as oportunidades e benefícios que essa nova tecnologia pode trazer. São, assim, identificados quatro tipos principais de risco:

- **Riscos Transversais**
- **Danos sociais**
- **Uso indevido**
- **Perda de controle**

Os riscos transversais referem-se a desafios técnicos e sociais que podem agravar os riscos específicos associados a essa tecnologia, e os riscos incluem dificuldades em projetar modelos seguros em domínios que são abertos (isto é, têm um amplo, aberto, campo espectro de aplicações), avaliar a segurança dos sistemas, rastrear seu uso, criarem-se padrões de segurança, lidar com viés, discriminação e desinformação gerada pelos modelos, garantir a

privacidade, confiabilidade e interpretabilidade, dentre muitos outros. Há ainda impactos socias, econômicos e uma grande incerteza sobre uma tecnologia ainda emergente e disruptiva, em que temos ainda muitas lacunas de conhecimento. Esses riscos destacam a importância de abordar questões éticas, legais e de segurança ao lidar com a IA de fronteira.

3.1 Riscos Transversais

Os riscos transversais envolvem desafios técnicos e sociais que permeiam todos os demais riscos e podem, por isso, potencializar e agravar os riscos específicos, como os sociais, relacionados a essa tecnologia. Muitos desses riscos advêm da dificuldade de se projetar modelos com fronteiras seguras em domínios tão abertos como os dos atuais modelos generativos, sendo difícil também avaliar a segurança desses sistemas. Se você tem escopo mais limitado, se as fronteiras do sistema com outros domínios são mais restritas, é mais fácil projetar os sistemas e sua segurança. Não parece, entretanto, ser o caso dos sistemas de IA que vem tendo aplicações nos mais diferentes campos e de diversas formas.

Esses sistemas ainda sofrem do “problema de especificação”. Você constrói uma arquitetura, mas exatamente o que o modelo faz não é especificado pelas empresas ou pelos responsáveis dos modelos, pois eles não são explicitamente programados (lembremos aqui da definição de Mitchell (1997) sobre aprendizado supervisionado que vimos antes). É assim, difícil prever o que o sistema vai fazer e, por conseguinte, sua segurança. Padrões de segurança ainda não estão estabelecidos e ainda há poucos incentivos que estimulem os criadores desses modelos a investirem em medidas de mitigação de risco.

Embora haja uma série de iniciativas para a criação de modelos abertos, há ainda uma concentração significativa do mercado de IA, e os altos custos iniciais associados ao treinamento dos modelos mais avançados parecem criar economias de escala e barreiras significativas à entrada de participantes menores. Por outro lado, o uso de sistemas abertos, ou mesmo o uso de APIs que conectam inúmeras aplicações a diferentes modelos de IA generativa, criam

uma grande exposição em vários pontos da cadeia desses produtos e que dificultam o controle e identificação dos acessos, a rastreabilidade, o uso indevido e o estabelecimento de responsabilidades por danos que possam ser causados por sistemas que fazem uso desses modelos.

Por último, embora possamos elencar muitos outros, há o viés inerente a esses modelos que, como vimos, basicamente reproduzem um padrão de dados já presente, e que potencializa o reforço de pré-conceitos e padrões institucionalizados mesmo que não eticamente adequados. Um exemplo simples, como o representado na figura 5, pode ser experimentado por qualquer pessoa com acesso à internet.

Figura 5. Imagens de pessoas fictícias produzidas por IA generativa através do site https://www.thispersondoesnotexist.com/. Em 10 imagens produzidas é claro o viés e o padrão das imagens criadas que exclui uma parcela grande da população de muitos países.

Riscos como esses devem ser um desafio nos próximos anos para a academia, pesquisadores e empresas, que são essenciais para a criação de modelos seguros e responsáveis, e para que se possa fazer um bom uso da tecnologia.

3.2 Danos Sociais

Existe uma gama de potenciais danos sociais decorrentes do uso direto e indireto dos modelos de IA generativa e da IA em geral. Dentre eles vamos destacar apenas três aqui:

a. *Degradação do ambiente de informação*. Seja pelo excesso de informações ou informações falsas. Isso pode levar os indivíduos a tomarem decisões perigosas e ainda reduzir a confiança em informações verdadeiras, aumentando a insensibilidade à informação. Em cenários de amplo alcance, isso pode contribuir para persistência de preconceitos sistêmicos, incentivar a polarização política e a violência, e exacerbar crises políticas e de saúde pública.

b. *Perturbação do mercado de trabalho*. Hatziuset al. (2023), economista chefe do Goldman Sachs, apresenta um estudo detalhado sobre os impactos da IA nos diferentes mercados de trabalho. Ele aponta para a IA substituindo o trabalho humano principalmente nas áreas jurídica e administrativa, embora com pouco efeito em trabalhos manuais e ao ar livre, como as áreas de construção, limpeza e extração. De qualquer modo ela aparece como ferramenta de produtividade em todos os outros campos, mais notadamente nas áreas financeira e de negócios, educação e tecnologia da informação. Essa transformação, tem o potencial de trazer mudanças disruptivas no mercado de trabalho com fortes impactos sociais.

c. *Preconceito, justiça e danos representacionais.* Os modelos podem conter e ampliar vieses já presentes nos dados de treinamento e que refletem desigualdades e estereótipos sociais históricos. Esses vieses, muitas vezes sutis, podem comprometer o uso equitativo e ético da IA, e a justiça das decisões que possam ter como base esses sistemas. Isso é particularmente preocupante em domínios do mundo real, como nas áreas de recrutamento para empregos, empréstimos financeiros, assistência médica e justiça, em que decisões tendenciosas podem reflexos significativos sobre a vida das pessoas.

3.3 Uso Indevido

Como muitas outras tecnologias, não obstante os benefícios que possa trazer, IA também tem o potencial de ser empregada com fins ilícitos, maliciosos,

e para trazer prejuízo a pessoas e organizações. Em resumo, pode ser empregada para o “mal”.

A capacidade de gerar conteúdo muitas vezes indistinguível da produção humana pode ser empregada para desinformação, violação de privacidade, manipulação e golpes, sendo possível a criação de *desinformação* direcionada*, personalizada*. Outros fins maliciosos de uso desses modelos são os riscos *cibernéticos,* que incluem a criação de automação de ataques cibernéticos, com riscos a infraestruturas críticas como energia, transporte, saúde e o sistema financeiro, que já são alvos frequentes desse tipo de ataque.

Por último, há ainda os chamados *riscos científicos de dupla utilização*. As tecnologias de IA, diferentemente de tecnologias como a nuclear e espacial, não possuem uso restrito, estão amplamente popularizadas e acessíveis, e podem ser usadas, por exemplo, para a criação de armas biológicas. É assim emblemático que muito recentemente um grupo de cientistas fizeram um abaixo-assinado contra armas biológicas geradas por inteligência artificial (ver, Estadão (2024)).

3.4 Perda de Controle

Modelos de IA, especialmente os de grande escala, podem se comportar de maneiras imprevisíveis, as chamadas “alucinações”, e a complexidade desses modelos dificulta a compreensão completa de seu funcionamento, aumentando o risco de perda de controle sobre suas ações.

Dois fatores parecem contribuir para potencializar esse risco de perda do controle. O primeiro, os humanos entregam cada vez mais o controle de decisões importantes, mesmo no seu dia a dia, às IA’s, o que torna cada vez mais difícil para os humanos retomar o controle. É o caso em que deixamos sistemas de recomendação “escolherem” por nós conteúdos que vamos ler ou ouvir, mas o risco é ainda maior quando entregamos nossas decisões a sistemas IA que podem estar desalinhados, como é o caso do potencial risco de acidentes em carros autônomos e talvez, no futuro, das aeronaves autônomas.

O segundo potencializador é o risco os sistemas de IA procurarem ativamente aumentar a sua própria influência e reduzir o controle humano. Embora pareça um cenário de ficção científica do tipo “as máquinas assumindo o controle”, o relatório do Departamento para Ciência, Inovação e Tecnologia do Reino Unido (Gov.uk, 2023) aponta que os sistemas de fronteira da IA já mostram sinais iniciais de capacidades que poderiam ser usadas para reduzir o controle humano. Mas nosso tempo e espaço aqui é reduzido e devemos deixar o aprofundamento de questões como essa para uma outra oportunidade.

4. Conclusão

O futuro da IA generativa promete ainda mais avanços, com modelos capazes de produzir não só conteúdos como vídeos e simulações complexas, mas estruturas e interações quase humanas. A integração dessas tecnologias em diferentes áreas pode transformar radicalmente como trabalhamos, como interagimos com a informação, e mesmo com a sociedade e outros sistemas. É, portanto, crucial que todos, não só pesquisadores, mas indivíduos e sociedade estejamos atentos aos seus desdobramentos.

Essa tecnologia tem o potencial de trazer muitos benefícios e se encontra em um momento de grande otimismo e inflação das expectativas. Mas existe uma incerteza significativa em torno de suas capacidades e riscos. Na linha das conclusões do relatório do Reino Unido (Gov.uk, 2023), é importante que os governos, o meio acadêmico, as empresas e a sociedade civil trabalhem em conjunto para enfrentar esses riscos, que são complexos e difíceis de prever, para mitigar os perigos potenciais e garantir que a IA beneficie a sociedade. Há o risco global da perda de confiança e fiabilidade nesta tecnologia, o que negaria permanentemente a nós e às gerações futuras os seus benefícios transformadores.

Por fim, deixo aqui uma mensagem para reflexão. Como vimos, a IA Generativa trabalha hoje basicamente com a linguagem, que é apenas uma representação que nós, seres humanos, fizemos da realidade. Algo muito

diferente, e ainda mais disruptivo, será quando a IA atuar, não mais somente sobre a linguagem, mas diretamente sobre as coisas e o mundo real.

Referências

BISHOP, Christopher M.; NASRABADI, Nasser M. *Pattern recognition and machine learning*. New York: Springer, 2006. Disponível em: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf. Acesso em: jul. 2024.

ESTADÃO. Cientistas fazem abaixo-assinado contra armas biológicas geradas por inteligência artificial. *Cultura Digital*, 2024. Disponível em: https://www.estadao.com.br/link/cultura-digital/cientistas-fazem-abaixo-assinado-contra-armas-biologicas-geradas-por-ia/. Acesso em: jul. 2024.

GARTNER. Beyond ChatGPT: The Future of Generative AI for Enterprises. 2023. Disponível em: https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises. Acesso em: jul. 2024.

GARTNER. What's New in Artificial Intelligence from the 2022 Gartner Hype Cycle. *Gartner*, 2022. Disponível em: https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2022-gartner-hype-cycle. Acesso em: jul. 2024.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep learning*. MIT Press, 2016. Disponível em: https://www.deeplearningbook.org/. Acesso em: jul. 2024.

GOV.UK. Frontier AI: capabilities and risks – discussion paper. From Department for Science, Innovation and Technology, 2023. Disponível em: https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper. Acesso em: jul. 2024.

HATZIUS, J.; BRIGGS, J.; KODNANI, D.; PIERDOMENICO, G. The Potentially Large Effects of Artificial Intelligence on Economic Growth Global Economics Analyst. *Goldman Sachs*, 2023. Disponível em: https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html. Acesso em: jul. 2024.

HUANG, S.; GRADY, P. Generative AI: A Creative New World. *Sequoia Capital US/Europe*, 2022. Disponível em: https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/. Acesso em: jul. 2024.

MIELKE, Sabrina J. et al. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. 2021. Disponível em: https://arxiv.org/abs/2112.10508. Acesso em: jul. 2024.

MITCHELL, T. M. Machine Learning. McGraw-Hill, Inc., 1997.

OLIVEIRA, R. IA Generativa: fundamentos, usos e riscos, 2024. In: Paisagens Híbridas – Cátedra Diseño, Arte y Ciencia 2024, Media Lab / MAUÁ – SP. Slides e link do youtube em:

https://github.com/Rogerio-mack/Catedra_Diseno_Arte_y_Ciencia_2024/blob/main/README.md.

PLAYGROUND TENSORFLOW. Tinker With a Neural Network Right Here in Your Browser. 2024. Disponível em: https://playground.tensorflow.org/. Acesso em: jul. 2024.

SCIKIT-LEARN. Scikit-learn Machine Learning in Python. 2024. Disponível em: https://scikit-learn.org/stable/. Acesso em: jul. 2024.

SEARLE, John R. Minds, brains, and programs. Behavioral and Brain Sciences, v. 3, n. 3, p. 417–424, 1980. ISSN 1469-1825. doi:10.1017/S0140525X00005756.

VASWANI, Ashish et al. Attention is all you need. Advances in Neural Information Processing Systems, v. 30, 2017.

ZHANG, Aston et al. Dive into deep learning. Cambridge University Press, 2023. Disponível em: https://d2l.ai/. Acesso em: jul. 2024.