6

COMPUTADORES NÃO ESTÃO NEM AÍ: UM POSSÍVEL LIMITE ÀS PRETENSÕES DE AUTONOMIA DA IA

Carlos Barth [1]

Computadores têm objetivos? Num sentido instrumental, parece razoável dizer que sim. Tome-se como exemplo o *Deep Blue*, sistema que venceu Kasparov no xadrez em 1997. Podemos encontrar ali uma cadeia de razões instrumentais: por que mover o peão? Para aumentar o domínio do centro do tabuleiro. E por que aumentar o domínio nesta região? Para movimentar as demais peças com mais segurança, e assim por diante. Essa cadeia de razões instrumentais pode ser rapidamente desenvolvida até chegarmos à razão maior dentro do jogo: vencer. Nesse sentido, podemos dizer que o *Deep Blue* está "tentando" ganhar o jogo. Vencer é, nesse cenário, uma razão última em função da qual seus demais objetivos se estruturam. Uma razão não instrumental desse tipo parece necessária para evitar um regresso sem termo de razões instrumentais. Contudo, é evidente que podemos ir além e perguntar: por que o *Deep Blue* tenta vencer o jogo? No caso de sistemas computacionais, a única resposta plausível é dizer que buscam vencer em função da forma como foram desenvolvidos. Não há alternativa exceto agir de modo conforme ao seu *design*.

É certo que seres humanos também estão sujeitos a esse tipo de imperativo. Nossas necessidades fisiológicas são o exemplo mais saliente: em dado momento sentiremos fome ou sede, e isso não depende de quaisquer outros objetivos que tenhamos. Mas podemos reconhecer algo parecido também no âmbito das nossas práticas

[1] Doutor em Filosofia pela UFMG. Bolsista de Pós-doutorado em Filosofia da FAJE.

culturalmente herdadas que se sedimentaram durante a ontogênese:[2] a forma como seguramos os talheres, a distância que mantemos uns dos outros em elevadores, o chiado ao pronunciarmos a letra "s" em certas palavras, e assim por diante. Não optamos por agir dessa forma em virtude de uma razão ulterior associada a nossos objetivos momentâneos. Antes, agimos dessa forma por sermos quem somos.[3]

Contudo, seres humanos têm a seu dispor uma fonte adicional de objetivos não instrumentais. Podemos imaginar, por exemplo, que as razões de Kasparov para querer vencer o jogo de xadrez não tinham origem na sua constituição. Talvez ele buscasse validar sua autoestima, ou a satisfação pessoal de obter reconhecimento público e admiração dos seus pares. O que distingue esse tipo de motivação é o envolvimento de uma autoimagem, uma interpretação de si mesmo como alguém capaz de bem desempenhar um determinado papel. Em síntese, de um *ego*.

De fato, o modo como agimos e como compreendemos o mundo não parece destacável do modo como compreendemos a nós mesmos. Um grande mestre do xadrez que perde o jogo para um iniciante pode se sentir envergonhado porque o modo como compreende a si mesmo não acomoda aquela possibilidade. Perder para um iniciante coloca em xeque o modo como ele é visto e compreendido, por si mesmo e pelos seus pares. Isso sugere que um episódio envolvendo vergonha não pode se dar como um fato isolado. Uma situação só emerge como vergonhosa na medida em que caracteriza uma ameaça a uma determinada interpretação de si mesmo.

---

[2] Para exemplos de como práticas culturais podem influenciar nossa constituição, ver Heyes (2018).

[3] Sobre práticas culturalmente sedimentadas, é tentador argumentar que seres humanos têm a capacidade de modificá-las. Isso é verdade, mas não é relevante para o ponto que se faz aqui. Podemos também alterar partes de nossa constituição biológica por vias medicinais, mas disso não se segue que características biológicas deixem de fazer parte do que nos constitui.

O que esse exemplo nos lembra é que existe uma dimensão afetiva envolvida no modo como os fatos do mundo aparecem para nós. Se alguma situação emerge como vergonhosa, celebrável ou preciosa é também porque ela se articula com nossa autoimagem. Esse profundo envolvimento de um *ego* na forma como compreendemos e lidamos com o mundo impõe um desafio para as pretensões da *inteligência artificial* (IA). Afinal, agir de modo inteligente pressupõe *inteligir* o mundo de um modo que torne aquele comportamento possível. Ao menos este é o argumento que John Haugeland nos apresentou ainda nos anos 1970 (Haugeland, 1998)[4].

O papel dessa dimensão afetiva é dado pela importância que damos ao que fazemos e ao que acontece conosco. Não somos indiferentes ao que acontece no mundo. Por isso podemos dizer que Kasparov sentiu o peso da derrota sofrida. Em contraste, a ausência dessa dimensão afetiva no *Deep Blue* nos permite dizer que, qualquer manifestação de alegria pela vitória (ou de tristeza por eventual derrota), não passaria de uma simulação vazia. Afinal, ele não estava realmente preocupado em vencer, apenas agia de modo conforme à sua constituição. Isso vale para o *Deep Blue*, mas vale também para computadores de forma geral. Vem daí a famosa síntese de Haugeland: "*o problema da inteligência artificial é que computadores não estão nem aí*"[5] (Haugeland, 1998, p. 47).

Mas em que medida a ausência dessa dimensão afetiva constitui um real problema para as pretensões da IA? No que se segue, buscaremos responder essa pergunta da seguinte forma: na seção II, contextualizaremos a questão no âmbito das pesquisas contemporâneas e mostraremos como o desafio de Haugeland pode apontar para uma

[4] Ver também Haugeland (1985).

[5] The trouble with artificial intelligence is that computers don't give a damn.

limitação concreta para as pretensões de autonomia da IA contemporânea, associada à capacidade de generalizar a partir do que nos é familiar. Na seção III, será analisada a forma como sistemas de IA generalizam a partir do que sabem, e as potenciais limitações dessa abordagem serão expostas nas seções IV e V.

II

Antes de mapearmos os potenciais efeitos do desafio feito por Haugeland, é preciso reposicioná-lo no contexto contemporâneo. Quando Haugeland o formulou, a IA era compreendida como uma empreitada científica. A tese predominante era a que Searle (1980) denominou "IA forte": um sistema computacional que replicasse corretamente os padrões associados aos processos cognitivos dos seres humanos estaria *efetivamente* pensando. Não precisaríamos nos limitar a dizer que um tal computador estaria simulando processos mentais, mas sim que ele *teria* uma mente dotada de inteligência. Cabia à IA, portanto, identificar esses padrões, expressá-los na forma de algoritmos e colocá-los para rodar em um computador. Nesse cenário, o argumento de Haugeland constitui um desafio significativo. Ele implica que o comportamento inteligente depende do envolvimento de um *ego*, mas ninguém fazia (nem faz), a menor ideia de como (ou se), é possível fazer com que um computador tenha um *ego*.

No entanto, contemporaneamente esse projeto da IA enquanto empreitada científica perdeu protagonismo. Em seu lugar, veio a IA como um projeto de engenharia, uma tecnologia utilizada para gerar aplicativos úteis. Vários desses aplicativos buscam substituir seres humanos na realização de certas tarefas, mas nenhum deles se preocupa em simular os mecanismos por meios dos quais a mente humana é capaz

de realizar essas tarefas. São, nesse sentido, como máquinas de lavar louça: podem realizar bem uma tarefa, talvez até melhor que nós mesmos, mas sem jogar qualquer luz sobre o modo como nós o fazemos. Uma vez que o foco está apenas no resultado obtido (louça limpa), não importa se o algoritmo utilizado é biológica ou psicologicamente plausível. Tais sistemas podem usar um poder de processamento e memória indisponível ao ser humano, por exemplo. Isso é aceitável, desde que o sistema alcance o comportamento desejado.

É verdade que não faltam pesquisadores afirmando que os sistemas de IA contemporâneos são, ou podem vir a ser, inteligentes. As discussões em torno de uma possível *inteligência artificial geral* (AGI, para *artificial general intelligence*), estão nessa órbita. Uma AGI seria uma forma de inteligência que, embora não necessariamente satisfaça os requisitos de uma IA forte, pode ser aplicada a um número ilimitado de domínios. Contudo, esta é uma posição bastante distinta daquela com que Haugeland operava. Ele estava interessado em compreender como era possível uma forma de inteligência humana, mas a ideia de uma AGI supõe que possam existir outras formas de inteligência. O abandono do ser humano como norte gera desafios particulares: como saber se estamos progredindo na direção certa? Precisamos de uma concepção "neutra" ou puramente funcional de inteligência, mas o número de concepções possíveis parece sempre próximo do número de pesquisadores engajados na empreitada de elaborá-la[6]. Assim, antes de verificarmos se Haugeland apresenta um problema para a possibilidade de uma AGI, precisaríamos ao menos concordar quanto ao que AGI significa.

À primeira vista, o desafio de Haugeland parece inócuo. Afinal, ele sugeriu que a dimensão afetiva é constitutiva do modo humano de gerar

[6] Em um levantamento, Hutter; Legg (2007) encontraram pelo menos 70 diferentes concepções.

comportamento inteligente, e essa só seria uma preocupação de quem está de fato tentando emular a mente humana em computadores. Mas há pelo menos uma pretensão da IA contemporânea que pode ser afetada pela ausência de uma dimensão afetiva: a busca por *autonomia*. Isto é, a criação de sistemas que possam tomar decisões independentes e de forma (supostamente) alinhada com nossos valores. Ao contrário do que muitas vezes se imagina, a dificuldade não está em elaborar uma formulação computacionalmente adequada do valor ou princípio que se deseja observar, mas sim em elaborar regras que guiem sua aplicação em situações concretas. Por exemplo, é relativamente fácil determinar que um sistema jamais minta, pois o "jamais" implica se abster de mentir em toda e qualquer situação, sem dar qualquer atenção às suas particularidades[7]. Contudo, é igualmente fácil formular exemplos em que essa regra tenha resultados indesejáveis: se instado a fazê-lo, deve o sistema revelar a um assassino onde está escondida sua potencial próxima vítima?

O que queremos, portanto, são sistemas que evitem mentir a menos que seja necessário. Porém, os fatores relevantes para determinar se estamos ou não diante de uma situação em que mentir é necessário variam enormemente caso a caso. No caso das IAs, podemos criar regras que estipulem o que fazer em situações conhecidas, mas o que dizer de situações inesperadas ou inéditas? Seres humanos têm a capacidade de reconhecer situações novas como semelhantes a situações com as quais ele já teve contato anterior, mas fazem isso sem perder de vista os elementos que tornam a situação corrente particularmente distinta. É por isso que recorremos com tanta

[7] O que se tem em mente aqui é um cenário em que o sistema deliberadamente opte for apresentar informações falsas em uma dada situação. Não confundir com o problema da alucinação (Ji *et al.*, 2022).

frequência a raciocínios como: “em geral devemos observar X, mas nesse caso concreto é preciso levar em conta que...”. Nesses casos, nós tendemos a gerar regras de improviso, adaptando o raciocínio de modo fluido e sensível ao que é circunstancialmente relevante. Mas o que nos permite fazer isso é, entre outras coisas, o fato de que nosso raciocínio é guiado por aquilo que *importa* pra nós. Essas regras geradas em tempo real tendem a ser adequadas porque são fruto do modo como interpretamos o mundo e a nós mesmos. Podemos colocar em dúvida, portanto, se e como essa formulação adaptativa, em tempo real, de regras adequadas a uma situação concreta pode ser emulada por sistemas que não dispõem da nossa dimensão afetiva, i.e. sistemas que não estão nem aí.

O que está em jogo, portanto, é o modo como generalizamos ou extrapolamos a partir do que nos é familiar. Com efeito, nosso próximo passo é compreender melhor como a capacidade de generalização se manifesta nos sistemas contemporâneos de IA.

III

A capacidade de generalizar (ou extrapolar) é essencial na IA contemporânea. Se os modelos gerados por *machine learning* fossem incapazes de generalizar para além das informações utilizadas durante o aprendizado, sua utilidade seria drasticamente reduzida. Não seria possível gerar sistemas de reconhecimento facial, por exemplo, mas tão somente sistemas que reconhecem o rosto das pessoas cujas fotos foram utilizadas para treinar o modelo (isto é, fotos presentes no *corpus* de treinamento). Mesmo essa capacidade seria severamente limitada, contudo. O sistema só conseguiria reconhecer um rosto em circunstâncias idênticas às que se apresentam no *corpus*, pois dados

fatores como incidência de luz, expressão facial e ângulo da foto, ele não conseguiria dizer se está a lidar com o mesmo rosto em circunstâncias diferentes ou com o rosto de outra pessoa[8].

Em modelos neurais, presentes na maior parte das pesquisas de IA contemporâneas, essa capacidade tem como base uma espécie de "tolerância ao erro". Dizer que um modelo aprendeu a reconhecer um rosto, por exemplo, significa dizer que o modelo aprendeu a associar uma determinada configuração de um conjunto de características (cor, forma, proporção, etc.), a uma certa categoria ("rosto do João"). Contudo, o modelo não exige que todas essas características estejam sempre presentes para que a categoria seja identificada. Basta que um conjunto suficientemente grande delas esteja presente, mesmo que possa haver certa variação no que é considerado suficiente, e no peso atribuído a cada característica. Assim, aquilo que permite ao modelo generalizar é também aquilo que faz com que ele possa errar ao realizar uma categorização, pois é possível que surja uma foto com o rosto do João em um contexto tão inesperado que o sistema não consiga reconhecer ali um conjunto suficiente de características conhecidas.

Com efeito, a capacidade de generalização das IAs contemporâneas é função da distribuição probabilística das propriedades presentes no *corpus* de treinamento. Se olhos e narizes aparecerem sempre (ou quase sempre), juntos, então o sistema aprenderá a supor que, onde quer que apareçam olhos, deve haver também um nariz. Embora isso permita feitos impressionantes, há potenciais limitações. Não por acaso, há uma subárea de pesquisa no interior da IA contemporânea chamada *out-of-distribution generalization* (algo como "generalização fora da

[8] Entre pesquisadores de IA, essa é uma situação conhecida como *overfitting*.

distribuição"), que se dedica a identificar e desenvolver outras formas de generalização.

Essa tem se mostrado uma empreitada notoriamente difícil. Vamos usar um exemplo para entendermos melhor a natureza da dificuldade. Deixemos o rosto de João de lado e suponhamos o desenvolvimento de uma IA que será utilizada para reconhecer a presença de elefantes em fotografias. O *corpus* de treinamento conterá, presumivelmente, inúmeras fotos de elefantes em diversas posições e ângulos. O objetivo do algoritmo de treinamento é identificar quais propriedades presentes no *corpus* podem ser associadas à elefantes e quais não podem. Isso significa que, se houver alguma propriedade comum à todas (ou quase todas) as fotos, a chance de o sistema tomar essa propriedade como sendo parte essencial de um elefante é grande. Suponhamos que todas as fotos tenham sido retiradas em dias ensolarados e sem nuvens. Isso faz com que o céu azul seja uma constante nessas fotos, gerando o risco de que a IA resultante busque pela presença abundante da cor azul quando estiver tentando determinar se há um elefante presente na foto ou não. Assim, bastará que apresentemos ao sistema uma foto de elefante feita durante a noite ou num dia nublado e o sistema não encontrará o que procura, gerando um falso negativo.

O algoritmo de aprendizado busca determinar o que é um elefante e o que não é a partir de características que ele encontra com suficiente frequência em fotos com elefantes. Se todas as fotos contendo elefantes contiverem também um céu azul, o algoritmo não conseguirá fazer o "recorte" adequado, e suporá que a cor azul é uma propriedade confiável na hora de detectar a presença de elefantes em fotos. Com efeito, se tudo o que o mecanismo de aprendizado tem à disposição para determinar a relevância de uma propriedade é a sua distribuição no *corpus* de treinamento, ele não conseguirá distinguir propriedades corretamente

mapeadas como pertencentes a elefantes e propriedades coincidentes, ou mesmos correlações espúrias.[9]

Podemos evitar esse tipo de problema? De um ponto de vista prático, sem dúvida. Basta acrescentar ao *corpus* de treinamento fotos de elefantes em ambientes mais diversos, tais como dias nublados, ambientes fechados e assim por diante. Isso não impede, contudo, que a dificuldade se manifeste de formas novas e inesperadas. Geirhos et al. (2019), por exemplo, apresenta um caso em que se percebeu uma tendência em atribuir um peso excessivo à textura da pele na hora de detectar elefantes, fazendo com que o sistema fique cego a casos em que a textura da pele é ocultada por uma sombra. Isso também pode ser corrigido, seja calibrando o algoritmo de treinamento para atribuir um peso menor a certas características, seja acrescentando dados que afetem a distribuição das propriedades presentes no *corpus*, fazendo com que o algoritmo as "enxergue" do modo desejado.

Contudo, a necessidade dessa ação humana para correção, e a impossibilidade de o sistema corrigir a si mesmo é justamente o problema. Tal necessidade emerge de limitações na capacidade dos sistemas de IA para extrapolar a novos tipos de situações. Por isso, tais sistemas ainda são profundamente dependentes de seres humanos que avaliam os resultados e identificam as situações no interior das quais o sistema tem condições de fornecer resultados confiáveis.

O mesmo raciocínio vale para sistemas que almejam interagir de modo direto e autônomo com seres humanos. Carros sem motorista

[9] Correlações espúrias são distribuições probabilísticas semelhantes que não implicam nenhum tipo de relação causal. O trabalho de Vigen (2015) é bem conhecido por apresentar exemplos curiosos, tais como a correlação entre o número de pessoas que morreram afogadas em piscinas e o número de filmes em que o ator Nicolas Cage aparece (para o período 1999-2009). Vale notar que detectar casos de correlação espúria pode ser particularmente difícil. Há, por exemplo, uma relação entre o tamanho da mão e a riqueza do vocabulário de uma pessoa, mas essa correlação não é espúria. Ela existe porque, em geral, crianças tendem a ter mãos e vocabulários menores do que adultos.

humano, robôs cirurgiões e *drones* militares são exemplos de aplicações contemporâneas da IA em que essa interação é bastante saliente. O *corpus* de treinamento dos modelos de IA utilizados nessas aplicações são, via de regra, registros de atividades humanas passadas. Porém, ao contrário da detecção de elefantes em fotos, estas são aplicações em que não temos controle sobre os contextos nos quais os sistemas de IA terão de tomar decisões. Em outras palavras, não temos como antecipar detalhadamente todos os possíveis cenários com os quais os sistemas terão de lidar. O cenário se agrava porque as consequências dessa limitação para aplicações como carros e armas autônomas são muito mais sensíveis do que um eventual falso negativo na hora de detectar um elefante em uma foto. A capacidade de generalizar com base em algo além da distribuição das propriedades parece essencial, portanto, para a geração de sistemas que possam atuar de forma autônoma. A seguir, vamos averiguar em maior profundidade os efeitos dessa distinção entre o modo como seres humanos e sistemas de IA generalizam.

IV

Na seção anterior, vimos que sistemas de IA contemporâneos generalizam a partir do que se mostra típico ou frequente. Com efeito, o manejo adequado de circunstâncias inéditas é um desafio para esses sistemas, uma vez que situações novas são, por definição, rompimentos com aquilo que é típico. Nem toda situação inédita é uma reconfiguração simples de situações familiares. Tal como durante uma investigação criminal, um único novo elemento pode levar a uma reinterpretação radical de todos os elementos com os quais já estávamos familiares: o principal suspeito pode se revelar inocente, um álibi pode se revelar

inválido, o relato de uma testemunha confiável pode se mostrar enganoso, e assim por diante.

Mas aqui podemos nos perguntar: quão comuns são exemplos de reviravoltas desse tipo? Isto é, com que frequência esse importar-se com o mundo realmente faz diferença em nossa atividade cognitiva? Se for algo relativamente raro ou presente apenas em um conjunto muito particular de decisões, então sua ausência não irá representar nenhuma grande preocupação para as pretensões de autonomia por parte da IA enquanto engenharia. Assim, ainda que se aceite a existência de diferentes formas de generalização, podemos nos questionar sobre o quão relevante ela é para os usos que pretendemos fazer dessa tecnologia. Em outras palavras, podemos nos perguntar em que medida essa capacidade de se importar com o mundo tal como ele é pode aparecer como um desafio para a IA enquanto engenharia.

O que vale para investigações criminais vale também para caracterizar situações simples e rotineiras no cotidiano. Decidir o que beber num jantar é sensível a um número indefinidamente multiplicável de fatores: quem está presente, a natureza da relação com quem está presente (é alguém confiável?), quem está ausente (há algum colega de trabalho por perto?), a temperatura ambiente, o horário, as condições de saúde de um parente distante, e assim por diante. Em síntese, não há como antecipar todos os fatores que precisarão ser levados em conta mesmo na hora de fazer escolhas simples como o que beber num restaurante[10].

Isso permite afirmar que há algo de inédito em *toda e qualquer situação.* Ainda que tenhamos o hábito de ir a um restaurante e pedir sempre os

[10] Um desenvolvimento mais detalhado desse caráter insaturável dos contextos de atividade humana pode ser encontrado em Barth (2024).

mesmos pratos, nunca deixamos de ser sensíveis a fatores específicos daquela ida ao restaurante, naquele dia, naquelas circunstâncias. Com efeito, a dimensão afetiva está presente na nossa lida com o mundo mesmo em situações repetitivas. Mesmo situações rotineiras são sempre únicas. De fato, se uma situação se repetisse em todas as suas particularidades isso rapidamente nos geraria estranheza, ou talvez uma sensação de *déjà vu*. Com efeito, há algo de único a cada situação, e nós sabemos lidar com essa particularidade de modo flexível, fluido, e sem perder de vista o que importa. Somos capazes de generalizar aplicações do nosso conhecimento para situações inéditas porque, dentre outras razões, o ineditismo das situações em que nos encontramos não faz com que percamos de vista aquilo que importa para nós.

Segue-se disso que encontrar uma forma de generalizar para além da distribuição de propriedades no *corpus* de treinamento é essencial para as pretensões de autonomia da IA. Evidentemente, não bastaria encontrar qualquer solução. É preciso encontrar alguma forma de generalização que seja equivalente à forma humana de generalizar a partir da dimensão afetiva. Assim, mesmo que a IA enquanto engenharia não tenha a pretensão de emular os mecanismos que caracterizam a cognição humana, na medida em que pretende introduzir no mundo sistemas artificiais que interagem autonomamente com seres humanos, ela precisa ir além da generalização via distribuição. Sem isso, ela não será capaz de reconhecer o que é relevante ou não em situações inéditas. O argumento de Haugeland permanece válido contra esta pretensão da IA contemporânea, portanto: o problema com os sistemas de IA contemporâneos que buscam atuar de forma autônoma é que eles não estão nem aí.

V

Caso estejamos no caminho certo, o cenário é bastante desafiador para as pretensões da IA contemporânea. Apesar disso, é preciso salientar que o argumento aqui apresentado é certamente insuficiente para uma afirmação categórica de que as pesquisas em IA jamais conseguirão superar essas limitações. A resposta mais honesta, portanto, é que simplesmente não temos como antecipar o que pode acontecer nos próximos anos. O que sabemos é que, até o momento, nenhum caminho realmente promissor foi revelado, e que o frequente otimismo encontrado entre pesquisadores de IA se ancora antes numa compreensão inadequada das limitações atuais.

Com efeito, é especialmente importante que sejamos capazes de analisar potenciais avanços. Assim poderemos saber se e quando estamos diante de um método capaz de modelar adequadamente a capacidade de generalização humana. Para um exemplo de como uma análise descuidada pode nos levar a conclusões precipitadas, vamos analisar aqui o caso dos grandes modelos linguísticos (*LLMs*, para *large language models*).

Num primeiro momento, parece difícil encontrar casos em que tais modelos cometam erros parecidos com os que vimos na seção anterior. Contudo, o poder das LLMs não vem do desenvolvimento de novas formas de generalização, mas sim do uso de uma quantidade gigantesca de informações no seu *corpus* de treinamento. Encontramos nele todo tipo de informação textual: postagens de redes sociais, artigos, livros, conteúdo de sites como a Wikipédia, e assim por diante. Por isso, a distribuição probabilística dos discursos presentes nesses materiais parece representar, em alguma medida, as decisões que tomamos, os comportamentos que adotamos e o modo como pensamos.

Considere, por exemplo, o modo como compreendemos relações físicas de causa e efeito. De modo geral, entendemos que as relações físicas que valem para um objeto podem ser generalizadas para outros objetos físicos: o que vale para carros, vale para motos e aviões. Mas um LLM não registra o mundo dessa forma. Ele não compreende o mundo físico em termos de relações físicas entre objetos e propriedades. Em vez disso, ele captura certos modos de articulação linguística que nós utilizamos para falar de objetos físicos. Ele consegue detectar, por exemplo, que certas construções linguísticas são comuns a carros, aviões, etc. Ao detectar que palavras como "velocidade" ou "colisão" aparecem associadas às palavras "carro" e "avião" no *corpus* de treinamento, o sistema pode trabalhar com a hipótese de que esses objetos têm algo em comum. Por essa via um tanto indireta, o sistema percebe que aquilo que pode ser dito de carros, pode também ser dito de motos e aviões.

No caso de LLMs, o que os algoritmos de aprendizado fazem é, portanto, reconhecer padrões no modo como nós expressamos nossa compreensão de mundo no passado. Ele consegue extrapolar, isto é, generalizar em alguma medida a partir destes padrões. Por isso, ele é capaz de fazer "palpites bem informados" sobre objetos e situações previamente desconhecidos. Mesmo que o *corpus* de treinamento não contenha informações sobre "espaçonaves", por exemplo, o LLM consegue inferir que certas palavras podem ser utilizadas para falar de espaçonaves (e.g. "decolar", "colidir"), enquanto outras são inadequadas (e.g. "beber", "comer"). Isso é o que está por trás das LLMs mais populares, como a família de modelos GPT, da OpenAI.

O exemplo das LLMs sugere que, mesmo sistemas incapazes de generalizar por outros meios que não a distribuição de propriedades é capaz de feitos impressionantes. Isso pode levar à adoção do seguinte

argumento: o que precisamos não são novas formas de generalização, mas sim *corpora* de treinamento cada vez mais adequados. Afinal, aquilo que importa pra nós está, de algum modo, codificado no conjunto de nossas decisões passadas sobre o que, como e quando falar ou escrever. Em outras palavras, tudo que precisamos é colocar a informação certa na distribuição certa. O grande sucesso das LLMs é uma evidência concreta do quanto esse caminho é promissor.

De fato, o sucesso desse tipo de sistema revela uma característica até então desconhecida da linguagem natural: é possível simular uma performance linguística plausível *sem* emular a forma como seres humanos compreendem linguagem. Este é um achado empírico significativo, pois revela algo sobre a linguagem que poderia ser considerado implausível há poucos anos. Mas não é a primeira vez que um nos deparamos com um achado desse tipo.

Quando Dreyfus (1972) apresentou sua famosa crítica aos anseios da IA clássica, era comum acreditar que grandes capacidades cognitivas, como a de jogar xadrez no nível de um mestre seria algo inalcançável a sistemas computacionais. As razões de Hubert Dreyfus têm como base as teses (fortemente inspiradas em Heidegger e Merleau-Ponty), que desenvolveu junto a seu irmão, Stuart Dreyfus sobre a *expertise* humana (Dreyfus; Dreyfus, 1986). Grosso modo, Dreyfus defende que nosso modo mais fundamental de lidar com o mundo tem um caráter utensiliar, não racional. Antes de sermos seres que calculam e fazem inferências lógicas, somos seres práticos. Não lidamos com o mundo como um matemático lida com números, mas sim como um jogador de futebol lida com a bola. Aprendemos a caminhar, a usar talheres, a demonstrar respeito. Aprendemos a forma adequada de nos portarmos em um número indefinidamente multiplicável de ambientes: jantares em restaurantes, emergências em hospitais, jantares em hospitais,

emergências em restaurantes, e assim por diante. Esse modo como seres humanos interagem com o mundo fundamenta a forma como esse mesmo mundo é compreendido.

Dreyfus era profundamente cético quanto à possibilidade de um computador demonstrar a capacidade de jogar xadrez com maestria porque acreditava que o modo como aprendemos a jogar xadrez é pautado por essa dimensão prática fundamental. Nós aprendemos a jogar xadrez não como aprendemos a fazer contas de matemática, mas sim como aprendemos a jogar futebol ou andar de bicicleta. Conforme praticamos, desenvolvemos uma capacidade cada vez mais refinada de perceber situações no xadrez como a de "ameaça à rainha" ou "área central desprotegida", tal como um jogador de futebol habilidoso percebe, de forma cada vez mais refinada, oportunidades de chute ou de drible. Temos essas habilidades ainda que não consigamos descrever tais situações em termos de condições necessárias e suficientes, mas a IA da época de Dreyfus exigia que todo conhecimento fosse disposto nessa forma. Assim, Dreyfus acreditava que a capacidade de jogar xadrez com maestria só estaria disponível a criaturas que interagisse com o mundo tal qual um ser humano interage.

O tempo desmentiu essa previsão. O *Deep Blue* não joga tal qual um ser humano jogaria. É um sistema especializadíssimo que é capaz de jogar xadrez com maestria, ainda que não apresente a relação prática e afetiva que humanos partilham com o mundo. Contudo, disso não se segue que a concepção de Dreyfus do ser humano como um ser fundamentalmente prático seja incorreta. O que isso nos diz é que há mais de um caminho possível para que um sistema, biológico ou não, possa demonstrar certa capacidade: com o *Deep Blue*, aprendemos que é possível jogar xadrez com maestria por um caminho distinto do humano.

O grande sucesso das LLMs em diversas tarefas envolvendo linguagem parece nos sugerir que algo semelhante pode ser dito acerca da linguagem. Assim como um dia foi surpreendente pensar que jogar xadrez com maestria é uma habilidade que poderia ser pensada de forma destacada da inteligência humana, há razões para pensarmos que o mesmo pode ser dito das nossas capacidades linguísticas (Mahowald *et al.*, 2023). Por isso podemos pedir ao ChatGPT que nos explique a moral de uma história, e a resposta dele pode ser surpreendentemente parecida, ou mesmo idêntica, à dada por criaturas linguísticas como nós. Isso pode ser feito mesmo que o LLM não consiga se colocar no lugar dos protagonistas, nem partilhe de seu embaraço ou alívio.

Ora, mas se é esse o caso, então talvez seja possível responder o desafio de Haugeland, ao menos nos domínios e tarefas em que a linguagem natural puder atuar como guia. A esperança é de que, com mais e mais dados, mais e mais contextos sejam capturados, a ponto de praticamente esgotar os contextos de atividade humana. Basta que se obtenha um volume suficientemente grande e adequadamente organizado para utilizar como *corpus* de treinamento. Assim, ainda que nós mesmos criemos novos contextos a partir da continua renovação de nossas formas de ser e agir no mundo, os padrões comportamentais resultantes poderão ser continuamente adicionados ao treinamento dos modelos. Basta que o sistema seja periódica ou continuamente retreinado com esses novos dados e, em tese, ele será capaz de emular a capacidade humana de se importar com o mundo a partir de uma codificação dos comportamentos resultantes dessa capacidade.

Otimistas geralmente se apegam a essa possibilidade, mas é importante reconhecer que ela caracteriza uma esperança, não uma tese bem fundamentada. Os limites das atuais abordagens são desconhecidos, e a história da IA está repleta de casos em que soluções

promissores se mostraram limitadas. Assim, não é nada claro que o desafio de Haugeland será superado apenas fazendo o que já estamos fazendo, só que mais rápido e com mais informações. Além disso, alguns pesquisadores já vêm notando sinais de esgotamento.

Considere, por exemplo, o estudo de Udandarao *et al.* (2024). Nele, os autores encontraram evidências de que a capacidade de codificar e acomodar cada vez mais contextos avança em escala logarítmica. Contudo, o acréscimo de informação necessário para permitir essa acomodação avança em escala exponencial. Isso significa que o volume de informações necessário para manter o ritmo de desenvolvimento é cada vez maior, e pode não haver informação suficiente no mundo para que essa abordagem cumpra sua promessa.

Caso isso se confirme, e supondo que nenhuma alternativa radicalmente distinta se mostre plausível, então o uso de tais sistemas continuará sempre dependente da avaliação de alguém que se importe com o mundo, isto é, nós mesmos. Essa é, portanto, uma limitação com grande potencial de frear os anseios de autonomia por parte da IA enquanto engenharia. Na mesma linha, ela impõe uma forte limitação à tese de que é possível "humanizar" a IA, no sentido de que podemos fazer com que ela atue de forma alinhada aos nossos anseios e valores (Christian, 2020).

VI

O cenário que se desenrola à nossa frente é incerto, escorregadio e mutável. Isso exige de nós um esforço constante para evitar conclusões precipitadas, independentemente de serem baseadas em uma visão otimista ou pessimista. A IA, em sua vasta capacidade de gerar fenômenos complexos, abre espaço para múltiplas interpretações,

tornando a análise da natureza de suas realizações uma tarefa desafiadora.

A habilidade humana de generalizar a partir de uma dimensão afetiva representa um obstáculo importante para os objetivos da IA contemporânea, na medida em que busca construir sistemas autônomos robustos. Aqui reside um ponto central no debate atual: até que ponto podemos confiar nas decisões autônomas que esses sistemas tomam? Os sucessos anteriores da IA não são garantias de êxitos futuros. Não podemos prever com precisão quando um sistema irá enfrentar uma situação que ultrapasse sua capacidade de generalizar adequadamente, o que pode resultar em pontos cegos na identificação de fatores cruciais para uma tomada de decisão eficaz.

Além disso, não há garantia de que essa limitação poderá ser completamente superada. O futuro dirá se estamos realmente alcançando um *plateau* no ritmo de evolução das capacidades dos sistemas de IA ou se estamos apenas passando por uma desaceleração momentânea, que poderá ser superada com o tempo e novas abordagens. Entretanto, se esse *plateau* se estender por muito tempo, podemos ser levados a reconsiderar a esperança de que é possível atingir a flexibilidade cognitiva humana por meio de tecnologias não humanas, pelo menos se continuarmos insistindo em uma abordagem centrada no acúmulo massivo de dados.

Caso sejamos forçados a abandonar a esperança nessa estratégia, é provável que haja uma sensação de desalento entre os pesquisadores envolvidos na empreitada. Porém, desconfio que os sistemas gerados a partir dessa tese não irão se importar.

Referências

BARTH, C. **Representational cognitive pluralism: towards a cognitive-science of relevance-sensitivity**. Tese de doutorado – Belo Horizonte: Faculdade de Filosofia e Ciências Humanas, Universidade Federal de Minas Gerais, 2024.

CHRISTIAN, B. **The Alignment Problem**. W. W. Norton & Company, 2020.

DREYFUS, H. **What computers can't do: a critique of artificial reason**. New York: Harper & Row, 1972.

DREYFUS, H.; DREYFUS, S. **Mind over Machine**: The Power of Human Intuition and Expertise in the Era of the Computer. 1986.

GEIRHOS, R. *et al*. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. **arXiV**, v. 1811.12231, 2019.

HAUGELAND, J. **Artificial Intelligence: The Very Idea**. Bradford, 1985.

HAUGELAND, J. Understanding Natural Language. Em: **Having thought**. Cambridge: Harvard University Press, 1998. p. 47–60.

HEYES, C. **Cognitive Gadgets**. Harvard University Press, 2018.

HUTTER, M.; LEGG, S. A Collection of Definitions of Intelligence. **arXiv**, v. 0706.3639, 2007.

JI, Z. *et al*. Survey of Hallucination in Natural Language Generation. **arXiv**, v. 2202.03629, 2022.

MAHOWALD, K. *et al*. Dissociating language and thought in large language models. **arXiv**, v. 2301.06627, 2023.

SEARLE, J. R. Minds, brains, and programs. **Behavioral and Brain Sciences**, v. 3, set. 1980.

UDANDARAO, V. et al. No Zero-Shot Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance. **arXiv**, v. 2404.04125, 2024.

VIGEN, T. **Spurious correlations**. New York: Hachette Books, 2015.