PRINCIPIA **27**(3): 609–623 (2023) doi: 10.5007/1808-1711.2023.e88209
Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

RESENHAS
REVIEWS

MITCHELL, Melanie. *Artificial Intelligence: A guide for thinking humans*. New York: Macmillan, 2019.

EROS MOREIRA DE CARVALHO. *Universidade Federal do Rio Grande do Sul,* BRASIL
eros.carvalho@ufrgs.br

RECEIVED: 29/05/2022 REVISED: 09/11/2022 ACCEPTED: 10/01/2023

1. Introdução

Desde que Alan Turing publicou o artigo "Podem os computadores digitais pensar?" (1951), essa pergunta tem inquietado não só pesquisadores da área de computação e da filosofia, mas também o público em geral. Será que os desenvolvimentos recentes e surpreendentes da inteligência artificial apoiam a crença de que em breve lidaremos com máquinas que pensam e são inteligentes? Em *Artificial intelligence: A guide for thinking humans* (2019), Melanie Mitchell, professora do *Santa Fe Institute*, nos EUA, nos convida para um mergulho no estado da arte da inteligência artificial para responder a essa questão. Se nos ativéssemos às manchetes sobre o tema, pareceria que as máquinas digitais já ultrapassaram e em muito as nossas habilidades intelectuais. Já foi anunciado que as máquinas ultrapassaram a capacidade humana de classificação de imagens, que podem ler um texto e responder questões sobre ele tão bem quanto uma pessoa, que são capazes de criar músicas belas, e que se saem melhor do que os mais experientes jogadores em jogos de estratégia, como o Go, inclusive realizando movimentos inusitados. Contudo, por mais que esses feitos sejam surpreendentes, eles em si mesmos não falam sobre como foram produzidos. E é na maneira como são produzidos que reside toda a diferença. Se não entendermos como as máquinas que estão por trás desses feitos funcionam, não poderemos afirmar apropriadamente se elas pensam ou não, se são inteligentes ou não. Tampouco poderemos fazê-lo se não adentramos na questão filosófica sobre o que é pensar e ser inteligente.

O livro tem a ambição de servir como uma introdução crítica às tecnologias mais recentemente desenvolvidas na área de inteligência artificial, tais como as redes neurais e a aprendizagem profunda. Mitchell não se limita a apresentar essas ferramentas em suas linhas gerais. Por meio de uma abordagem passo a passo, ela progressivamente nos mostra como elas estão estruturadas e funcionam. O nível de detalhe das

descrições e explicações do funcionamento dessas tecnologias é muito bem balanceado, pois tanto não sobrecarrega o leigo com terminologias herméticas quanto não desmotiva o profissional bem informado da área. Enquanto percorre essas tecnologias e as maneiras pelas quais elas têm sido desenvolvidas, aperfeiçoadas e aplicadas no nosso dia a dia, Mitchell constantemente se pergunta também se elas estariam nos deixando mais próximos da construção de uma máquina que pensa. Nesses momentos, entramos em terreno filosófico. Embora a autora não seja uma filósofa, ela tem um faro filosófico aguçado. Sem muito jargão filosófico, ela esbarra e trata com propriedade questões éticas e metafísicas suscitadas pela IA. Trata-se, portanto, de um livro que contempla os interesses de um público muito vasto, desde leigos até profissionais da computação e das humanidades interessados em conhecer o estado da arte da inteligência artificial, discutir a questão de Turing à luz dos avanços mais recentes da IA e pensar as consequências éticas e sociais da presença massiva da inteligência artificial em nossas vidas. O livro é muito bem-sucedido em vários aspectos.

Ao longo do livro, e na contramão da euforia recente, Melanie Mitchell deixa bastante claro o seu ceticismo em relação à inteligência artificial no que diz respeito à ambição de construir máquinas inteligentes. Embora ela reconheça que as tecnologias correntes possibilitaram e possibilitarão feitos extraordinários em tarefas como reconhecimento de imagens, processamento de linguagem natural, diagnóstico de doenças etc., as principais dificuldades e questões teóricas enfrentadas pela inteligência artificial nos anos 70 e 80 permanecem em aberto. Por mais eficientes que sejam as máquinas de hoje, elas não têm qualquer entendimento do que estão fazendo e, o mais alarmante, elas não são confiáveis e robustas. Logo adiante explicarei por quê.

O livro está dividido em cinco partes e contém dezesseis capítulos. Na primeira parte, "Background", que compreende os três primeiros capítulos, a autora oferece um panorama e uma breve história da disciplina da Inteligência Artificial. Na segunda parte, "Looking and Seeing", que compreende os 4 capítulos seguintes, aprendemos sobre as redes neurais, a aprendizagem profunda e o relativo sucesso que essas tecnologias obtiveram na tarefa de reconhecimento de imagens. Na terceira parte, "Learning to play", que compreende os capítulos 8 até 10, Mitchell explica como essas tecnologias foram aprimoradas para serem aplicadas a jogos de estratégia. Na quarta parte, "Artificial Intelligence Meets Natural Language", que compreende os três capítulos seguintes, aprendemos como as redes neurais foram aplicadas a tarefas que envolvem o processamento de linguagem natural, tais como o reconhecimento da fala e a tradução. Por fim, na última parte, "The Barrier of Meaning", que compreende os três últimos capítulos, Melanie Mitchell apresenta e explora as suas principais reservas em relação à IA, em especial o seu ceticismo quanto à capacidade das máquinas de compreender o que estão fazendo, de abstrair e de adquirir capacidades gerais que possam ser aplicadas não apenas às tarefas mais específicas em que foram treinadas.

2. AI simbólica e subsimbólica

Começo introduzindo duas distinções que nos ajudam a compreender o lugar da obra de Mitchell na história da inteligência artificial e no debate da filosofia da inteligência artificial. A primeira distinção tem a ver com dois interesses que sempre estiverem presentes na pesquisa sobre a inteligência artificial. Um deles é o interesse teórico, que consiste em usar ferramentas computacionais para tentar emular e compreender processos cognitivos em geral. É nesse contexto que a inteligência artificial entrou como uma disciplina chave na formação das chamadas *ciências cognitivas* e fomentou a metáfora da mente como uma espécie de computador. O segundo interesse é prático e tecnológico, ele consiste em usar ferramentas computacionais para resolver problemas de modo automático e eficiente, sem qualquer preocupação em emular a maneira como nós humanos resolvemos ou resolveríamos esses problemas. Esses interesses não são excludentes e sempre estiveram presentes na área de IA. O foco do livro de Mitchell é sobretudo a pesquisa em IA tomada a partir do interesse teórico. Como já mencionado, a sua questão de fundo é se a área avançou na emulação do pensamento ou da inteligência, embora ela não negligencie as consequências éticas do emprego da IA mesmo enquanto mera tecnologia e há um capítulo no livro dedicado ao assunto.

Uma segunda distinção, a distinção entre IA simbólica e subsimbólica, é relevante para entender por que há uma frenesi nos dias de hoje em relação à IA. Quando a inteligência artificial surgiu, nos anos 50 e 60, ela era dominantemente simbólica. Isto quer dizer basicamente que as arquiteturas computacionais propostas inicialmente se inspiraram no nosso conhecimento explícito, que é articulado linguisticamente. Isso não significa obviamente que essas máquinas tivessem já que "falar" português ou inglês, mas que os dados com os quais elas operavam tinham de estar estruturados simbolicamente para possibilitar o emprego de uma máquina automática de inferências dedutivas e até probabilísticas. O maior sucesso desse tipo de IA foram os sistemas especialistas. Trata-se de sistemas que visam codificar em regras explícitas o conhecimento que especialistas têm sobre um determinado assunto. Por exemplo, o sistema de IA DeepMind[1] codifica o conhecimento de oftalmologistas sobre doenças dos olhos e funciona como um sistema auxiliar de diagnóstico. Contudo, nos anos 80, ficou claro que esse tipo de abordagem não poderia ser a chave para a inteligência em geral. O filósofo Hubert Dreyfus, autor do livro *What Computers Can't Do* (1979), foi uma figura muito importante neste período para trazer à tona as limitações da abordagem simbólica. Logo se percebeu, como foi bem capturado numa frase célebre de um dos fundadores da IA, Marvin Minsky, que as coisas mais fáceis são as mais difíceis. Por exemplo, é muito fácil para nós reconhecer objetos e navegar no ambiente ao nosso redor, mas é extremamente difícil, talvez impossível, programar máquinas para fazê-lo a partir da abordagem simbólica. Tampouco parece ser o caso

que aprendemos a discriminar e a reconhecer diferentes tipos de objetos com base em regras explícitas. Talvez nem todo conhecimento seja explicitável e, se assim for, outras abordagens são necessárias se queremos emular em uma máquina o comportamento inteligente. Essas dificuldades levaram a IA a enfrentar o seu primeiro inverno no final dos anos 80 e ao longo dos anos 90. Os recursos minguaram e o interesse pela área feneceu.

Neste cenário, a atenção dos pesquisadores em IA começou a se voltar para a abordagem subsimbólica. Essa abordagem não surgiu após a derrocada da abordagem simbólica. Ela esteve presente desde o início da IA, mas foi eclipsada pelo entusiamo inicial com a abordagem simbólica e o sucesso dos sistemas especialistas e provadores de teoremas matemáticos que esta última propiciou. Inspirada em ideias que emergiram nos anos 30 e 40, na cibernética, tais como a causação circular, a importância da retropropagação para o controle do comportamento e o fenômeno da auto-organização biológica (Boden 2020, p. 28-32), a abordagem subsimbólica busca emular a auto-organização exibida pelos seres vivos e, em particular, o funcionamento do cérebro. As redes neurais são uma das principais apostas desta abordagem para produzir comportamento inteligente. Nas duas últimas décadas, as redes neurais foram aperfeiçoadas por técnicas de aprendizagem de máquina, que são técnicas que possibilitam que a máquina aprenda através da própria "experiência". Uma técnica de aprendizagem em particular obteve resultados surpreendentes e passou a receber muita atenção na comunidade de IA, a aprendizagem profunda. Desde então, a área tem experimentado uma nova primavera.

3. Redes neurais e aprendizagem profunda

Embora a aprendizagem profunda (*deep learning*) seja apenas uma técnica na subdisciplina de aprendizagem de máquina (*machine learning*), ela se tornou a técnica dominante e a própria IA, indesejavelmente, segundo Mitchell, é vista atualmente quase como sinônima de aprendizagem profunda. Essa técnica, como dito, é aplicada às redes neurais e, por isso, vamos começar com elas. Farei uma apresentação bem geral, seguindo a autora na suas primeiras formulações da técnica. Ao longo do livro, ao explicar como a técnica é aplicada a problemas concretos de processamento de imagens, voz, linguagem etc., a autora discute ajustes e variações dessa técnica em mais detalhes.

Redes neurais são arquiteturas computacionais que recebem esse nome por se inspirarem no funcionamento dos neurônios. A ideia básica é que elas são compostas por unidades de processamento de informação conectadas entre si, formando uma rede. Essas unidades têm entradas de dados e uma saída. Além disso, elas também possuem uma variável peso, para indicar a força das suas conexões com outras uni-

dades, e uma variável limiar, para indicar quando emitir o valor de saída. Assim, uma unidade recebe dados que são então multiplicados pelo seu peso e posteriormente somados. Se o valor resultante superar o limiar, então a unidade emite um sinal de saída.

Suponha, seguindo um exemplo oferecido por Mitchell, que queiramos construir uma rede neural capaz de reconhecer dígitos. Os dígitos são fornecidos em imagens de 16 vs. 16 pixels, totalizando 256 pixels. Os pixels de uma imagem constituem os dados de entrada da nossa rede neural. Esses dados alimentam aquilo que se chama de uma camada escondida da rede neural. As camadas escondidas compreendem as camadas entre a entrada e a saída da rede. Uma camada é formada por várias unidades. No caso da primeira camada escondida, cada pixel da imagem de entrada é conectado a cada uma das suas unidades. Assim, a unidade 1 recebe 256 dados e o mesmo se aplica às demais unidades. No caso em tela, as imagens que servirão de entrada para a rede são todas acinzentadas. Assim, cada um dos dados é um valor de 0 a 1 onde 0 representa branco, 1 preto e os valores intermediários representam o grau de acinzentado. Ao entrar na unidade, esses dados são multiplicados pelo peso da unidade e em seguida somados. Se ultrapassarem o limiar da unidade, o resultado é passado adiante. O mesmo acontece com a unidade 2 e todas as demais unidades dessa primeira camada escondida. Como as unidades obtêm o valor do peso e o valor do limiar? Inicialmente, esses valores podem ser distribuídos de modo aleatório. O que nos interessará depois é saber como eles são atualizados. Na sequência, cada unidade da primeira camada escondida está conectada a cada unidade da camada seguinte, que também é composta por um conjunto de unidades. No caso suposto, a rede neural tem apenas duas camadas. Então a segunda camada é a camada de saída. Como o objetivo é que ela reconheça dígitos, vamos supor, decimais, então a última camada será composta por dez unidades, cada uma representando um dos dígitos de 0 até 9. Essas unidades também têm peso, que serão usados para multiplicar os dados recebidos das unidades da camada escondida anterior. Esses resultados são somados e o resultado é liberado pela saída da unidade. Como se trata de uma camada de saída, as unidades que a compõem não têm um limiar de saída. Qualquer que seja o valor, ele será liberado. Esse valor deve ser lido como indicando a confiança de que a imagem contém o dígito que aquela unidade representa. Assim, se a unidade 1 da camada da saída contém o valor 0.8, e ela representa o dígito 0, então isso significaria que há 80% de chance de que o dígito na imagem de entrada seja o dígito 0.

Não é de se esperar que, de início, os resultados sejam corretos ou promissores. Por isso as redes neurais precisam ser treinadas, e aqui entra a aprendizagem profunda. É preciso calibrar os valores dos pesos das unidades de todas as camadas e os valores dos limiares das unidades de todas as camadas escondidas, que, no exemplo trabalhado, é uma só. A atualização desses valores é realizada por um algoritmo de retropropagação. Suponha que a unidade 1 da camada de saída indica 90% de

chance que o dígito da imagem seja 0, mas o dígito da imagem é 9. Esse é um péssimo resultado. O algoritmo propaga esse erro para que as unidades das camadas antecedentes ajustem os seus pesos e limiares. Quanto maior a magnitude do erro, maior a magnitude do ajuste. O mesmo ocorre em relação às demais unidades de saída. Se alguma delas acerta, a propagação de erro não ocorre nesse caso. Repare que, para esse algoritmo funcionar, é preciso que tenhamos independentemente a informação correta sobre qual é o dígito representado na imagem. Assim, para treinar uma rede neural, precisamos de imagens que já estejam catalogadas. O conjunto dessas imagens é normalmente chamado de *conjunto de treinamento*. A ideia é que por meio de sucessivos ajustes dos pesos e limiares das unidades da rede, ela convirja para valores que respondam de modo confiável ao tipo de imagem visado. No exemplo dado pela autora, a rede de duas camadas para a detecção de dígitos precisou ser treinada com 60 mil imagens de dígitos. Após o treinamento, um segundo conjunto de imagens, o *conjunto teste*, é usado para aferir a confiabilidade da rede neural. Suspende-se o algoritmo de retropropagação e computa-se os acertos e erros de classificação da rede neural ao ser alimentada com as imagens do conjunto teste.

As redes neurais podem ter n camadas escondidas. E as camadas podem ter m unidades. Como dimensionar a rede é algo que depende da tarefa em questão e varia muito de uma para outra. Por exemplo, uma rede para detectar imagens de cachorros e gatos normalmente tem 4 ou mais camadas. Como Melanie Mithcell comenta, determinar esses parâmetros ainda se trata de uma arte e depende de muita tentativa e erro. O termo *aprendizagem profunda* vem do fato de que muitas dessas redes neurais têm dezenas de camadas. E são redes que aprendem automaticamente, desde que contem com dados já classificados. Embora bastante simples, na sua estrutura mais geral, essa técnica é capaz de encontrar soluções para uma vasta classe de problemas. Essa é a técnica que está na base da maioria das máquinas que hoje realizam reconhecimento de imagens e voz, processamento de linguagem natural, que possibilitam tecnologias como a Alexa e o Siri, e que jogam jogos de estratégia, embora, em cada caso, inovações e adaptações tiveram que ser engendradas para estender o domínio de aplicação da aprendizagem profunda. Mitchell descreve cada um desses casos com razoável detalhe no livro.

4. São confiáveis?

Essas máquinas são confiáveis? Uma boa maneira de estimular e promover a busca por inteligência artificial cada vez mais eficaz na resolução de uma tarefa é por meio de uma competição. Na área de reconhecimento de imagens, a ImageNet[2] cumpriu essa função na última década. Nesta competição, os participantes recebem o mesmo conjunto de imagens de treinamento. É uma base gigante, que passa de um milhão

de imagens, de diversas categorias. Durante a competição, para cada imagem do conjunto de teste, as redes neurais podem fazer até cinco sugestões de classificação. Na primeira edição, a rede neural campeã acertou 72% dos casos. Em 2012, a campeã chegou à surpreendente marca de 85% de acerto. Em 2015, já beirava os 98%, o que levou alguns a afirmar que a IA tinha ultrapassado os humanos no reconhecimento de imagens.

Será mesmo? Como argumenta Mitchell, depende. Em primeiro lugar, a taxa de sucesso de 98% refere-se ao sucesso no teste em que a máquina pode fazer 5 sugestões para cada imagem. Quando o teste é restrito a uma única sugestão, a taxa de sucesso cai para 82%. O resultado ainda é impressionante, mas não o suficiente para dizer que ultrapassou a capacidade humana de reconhecimento de imagens. Em segundo lugar, as redes neurais acertam mais quando os objetos alvo estão sozinhos ou recebem foco, em contraste com o pano de fundo mais difuso. Se há mais de um objeto, ou o objeto alvo é muito pequeno, ou está desfocado, as redes neurais tendem a errar. Nós não temos tanta dificuldade de identificar um cachorro numa imagem mesmo que ele seja um entre vários objetos e ocupe uma parcela pequena da imagem. Melanie Mitchell conta o caso de uma rede que foi treinada para identificar cachorros e que, por acaso, foi submetida durante o treinamento a um número significativo de imagens de cachorros onde havia também um frisbee no cenário. O resultado é que a rede passou a classificar imagens de frisbee também como imagens de cachorro. Assim, não é claro o que exatamente essas redes estão aprendendo. Não parece que elas estejam de fato identificando um objeto específico numa certa região da imagem, mas capturando um padrão geral de pixels que, embora não seja específico de cachorros, pode ser eficaz para detectar várias fotos de cachorros, mas não apenas. Se for isso, então essas máquinas não têm o que os filósofos chamam de *intencionalidade intrínseca* (Adams & Aizawa 2001, p. 48), a capacidade de dirigir-se a algo específico no mundo. Em virtude dessas lacunas e falhas, as competições mais recentes exigem que a localização do objeto alvo na imagem também seja identificada. Os resultados continuam bons, como relata Mitchell, mas não tão surpreendentes quanto antes. Em terceiro lugar, e o que é mais grave, é muito fácil produzir imagens que enganem facilmente uma rede neural já bem treinada. Tanto é possível pegar uma imagem de cachorro e transformá-la numa imagem que, para nós, é apenas ruído, mas que ainda é identificada pela rede como uma imagem de cachorro, como é possível fazer pequenas alterações nessa imagem que são imperceptíveis para nós, mas que a rede já não classifica mais como uma imagem de cachorro. Isso mostra que a capacidade dessas redes de reconhecer imagens não é robusta, pelo menos não tanto quanto a nossa, e que a sua eficácia é restrita a imagens que atendem requisitos bem estritos.

Em virtude desses erros grosseiros, Mitchell conclui que as redes neurais não têm qualquer compreensão do que estão identificando e reconhecendo. Não precisamos apelar para a atividade complexa de interpretar e descrever a situação que uma ima-

gem pode representar, como Mitchell faz em dois momentos do livro — a foto de uma soldada voltando da guerra e encontrando no aeroporto o seu cachorro que ela não via há tempos e a foto de Obama pregando uma peça em um amigo que acaba de subir em uma balança. Sem que este último perceba, Obama coloca o pé na balança, vários conhecidos ao redor notam e estão a rir da situação — , para sustentar que as máquinas não têm o entendimento que temos, já é suficiente que elas não sejam robustas na identificação de objetos simples, isto é, não sejam capazes de identificar um tipo de objeto em diferentes contextos e situações. Outro teórico da IA, Judea Pearl, sustenta que a causa para esses erros grosseiros é que as redes neurais estão rastreando meras correlações, e que elas precisariam ser equipadas com conhecimento e raciocínio causal. Em uma entrevista recente, ele afirmou que "todas as realizações impressionantes da aprendizagem profunda não passam de ajuste de curva" (2018). Esse comentário é interessante, pois assinala como o problema clássico da indução é crucial para a IA. As redes neurais não rastreiam relações causais, elas não distinguem boas de más induções.

Como aprendemos ao longo do livro, a falta de robustez é um traço de todas as tecnologias que se baseiam em redes neurais treinadas por aprendizagem profunda. Mesmo quando saímos do âmbito do reconhecimento de imagens e entramos no âmbito do processamento de linguagem natural, tradução etc., também essas redes não são robustas e podem facilmente ser enganadas. Embora elas sejam eficientes em contextos bem específicos, e são bem úteis quando usadas nesses contextos, elas falham muito quando são aplicadas fora desses contextos. No caso de processamento de linguagem natural, por exemplo, as redes são mais eficientes quando lidam com frases curtas do que longas. O gargalo neste domínio é ainda maior do que no reconhecimento de imagens, pois atividades como o de tradução e compreensão de texto são ainda mais dependentes de conhecimento de fundo. Em todo caso, o ponto crucial é que as "capacidades" adquiridas pelas redes neurais não são generalizáveis e flexíveis. O Alpha Go, que fez tanto sucesso alguns anos atrás ao vencer o campeão mundial de Go, não faz nada mais além de jogar Go. A rede por trás do Alpha Go precisaria ser reconfigurada e retreinada para jogar outro jogo, e então não seria mais capaz de jogar Go. As nossas habilidades podem não ser completamente gerais, mas elas certamente não são tão estreitas, e admitem uma flexibilidade que não encontramos nessas máquinas. A marca da perícia, mesmo a perícia motora (e.g. acertar o pênalti ou a cesta de basquete), não é a repetição, mas a variação, isto é, ser capaz de atingir o objetivo de diferentes maneiras.

Outra diferença crucial: nós aprendemos com alguns poucos exemplos a discriminar certos tipos de objetos. Como vimos, essas redes precisam ser treinadas com milhares de exemplos. Nesse sentido, as redes neurais são ainda mais dependentes do que nós do conhecimento (humano) dos outros, já que o seu treinamento requer dados já classificados. Se tomamos o caso das imagens, de onde vêm as imagens já

classificadas para o treinamento de redes neurais? Elas vêm de gigantescos bancos de dados que foram criados justamente para alimentar redes neurais. Provavelmente, sem que você saiba, você já contribuiu para a formação desses bancos de dados. Toda vez que você reconhece imagens em um site para certificar que você não é um robô, você está ajudando na construção de alguma base de dados. Mas a maior parte do trabalho vem de empresas, como a Mechanical Turk[3], da Amazon, que contratam mão de obra barata em países periféricos para realizar essa tarefa.

Por fim, também não temos qualquer razão para pensar que por trás da nossa capacidade de reconhecer imagens e lidar com a linguagem esteja uma rede similar a esta que descrevemos. As redes neurais computacionais são apenas levemente inspiradas no funcionamento do cérebro. Em primeiro lugar, as redes neurais têm um funcionamento linear. Dados entram, são processados e, durante a fase de treinamento, o sinal de saída gera um efeito de retropropagação para a calibragem das unidades da rede. O nosso cérebro funciona de maneira dinâmica o tempo inteiro, entradas perceptuais influenciam saídas motoras e vice-versa simultaneamente. As conexões são muito mais cruzadas, não há divisão em camadas. Em segundo lugar, as redes neurais, quando muito, são um modelo apenas da dimensão elétrica do funcionamento do cérebro. No entanto, as reações químicas entre neurotransmissores, as trocas de elementos mediadas pela circulação sanguínea, e mesmo as interações metabólicas e mecânicas em virtude da sua estreita conexão com o corpo são elementos essenciais para o funcionamento adequado do cérebro. A nossa mente, mesmo que seja superveniente ao cérebro, o que é questionado por abordagens corporificadas e estendidas da mente, não é superveniente apenas a sua atividade elétrica. Se a chave para emular a mente é emular o cérebro, essas redes neurais ainda estão muito longe de fazê-lo.

5. O teste de Turing

Turing propôs o jogo da imitação como um substituto para a pergunta "podem máquinas pensar?" (1950). Há várias versões do jogo da imitação. Em uma delas, um interrogador deve decidir, após cinco minutos de conversa por meio de chat com A e B, qual dos dois é um humano e qual deles é um computador. O interrogador não pode ver ou ouvir ou ter qualquer contato perceptual com A ou B. Para Turing, essa restrição era importante para que o teste rastreasse apenas as capacidades intelectuais humanas. Assim, um robô humanoide aparentemente indistinguível de um humano em uma conversa face a face não atenderia o seu propósito. Sua previsão, em 1950, era a de que em 50 anos teríamos máquinas que passariam no seu teste, isto é, que enganariam um interrogador mediano em pelo menos 30% dos casos.

Como vimos, nada como uma competição para fomentar a área de IA. Há com-

petições periódicas em torno do teste de Turing. Em 2014, o Eugene Goostman foi o primeiro chatbot a passar no teste de Turing, em uma competição organizada pela Royal Society (Warwick & Shah 2016). Tanto o Eugene quanto os chatbots que fizeram sucesso nos anos 60 e 70, como o Eliza (1964)[4] e o Parry (1972)[5], foram construídos com base na abordagem simbólica, enquanto os mais recentes, como os embutidos no Siri e Alexa, baseiam-se em redes neurais.

Mas será o teste de Turing um bom teste para a inteligência? A crítica mais comum, endossada por Mitchell, é que estas máquinas estão apenas imitando o comportamento humano. Algo não inteligente pode causar um comportamento que parece inteligente. Uma reação possível é propor testes ainda mais exigentes, como relata Mitchell no livro. Em uma versão mais severa do teste, temos três interrogadores, que devem conversar por duas horas com três humanos e um computador, sem saber quem é quem. Ao final, cada interrogador deve dar a cada um dos quatro o veredito "humano" ou "máquina". Além disso, cada interrogador deve ranquear os entrevistados como mais ou menos humano. A máquina passará no teste se enganar dois ou mais interrogadores e se a média de ranqueamento do computador for igual ou maior do que a média de ranqueamento de dois ou mais dos humanos entrevistados. Segundo a autora, não há notícia de que alguma máquina tenha passado neste teste mais severo. Há inclusive apostas[6] sobre quando esse feito será obtido.

Na minha avaliação, não parece que esse teste mais severo altere muito o que podemos extrair dele. Talvez até se possa dizer que as chances de que ele rastreie a inteligência ou o pensamento são maiores, em comparação com o teste original, mas permanece aberta a possibilidade de máquinas que não têm inteligência alguma estejam apenas imitando comportamento inteligente. O teste em si mesmo revela mais sobre como nós podemos ser enganados do que sobre a inteligência. Sem uma discussão sobre o que é a inteligência e como ela é e pode ser gerada, não há como contornar essa lacuna. Além disso, o teste é limitado por se concentrar numa atividade muito específica, a conversação. A inteligência humana se manifesta de muitas maneiras. É difícil não ver a lida habilidosa de esportistas e dançarinos experientes e peritos como manifestação de inteligência. Mas ao exigir que o corpo fique às escondidas no teste, Turing deixou de lado esse tipo de inteligência corporal para se concentrar nas capacidades intelectuais humanas (1950, p. 434), como se elas fossem internas e independentes do corpo e das nossas habilidades corporais em geral. Eu arriscaria dizer que o teste de Turing contrabandeia o comprometimento com uma versão materialista do mito do fantasma na máquina: o mito da máquina (o computador) dentro da máquina (o corpo).

Isso não significa que o teste de Turing não possa ser reelaborado de maneiras interessantes, especialmente se associado a uma discussão sobre a natureza da inteligência. Na literatura especializada, encontramos várias propostas de aperfeiçoamento do teste de Turing nessa direção. Uma delas é que se leve em consideração

também o juízo dos competidores sobre se os interrogadores são humanos ou não. O objetivo é ter um teste mais simétrico capaz de capturar o reconhecimento mútuo, elemento fundamental de uma concepção social da inteligência (Mallory 2020). Outro teste forjado para evitar o risco de que algo não inteligente simule uma conversa inteligente é o teste do questionamento. O objetivo é que o competidor produza um questionário sobre um tema que será, então, avaliado pelo interrogador quanto a sua correção e qualidade. Por exemplo, em um questionário de Detetive, o competidor deve interrogar o suspeito, e seu interrogatório será avaliado por um detetive especialista (Damassino 2020). A ideia subjacente é que é muito improvável que um agente produza um interrogatório competente acerca de um tema sem compreender esse tema. Nada disso aparece no livro de Mitchell, que não tem a pretensão de revisar a já longa e complexa discussão em torno do teste de Turing (veja, por exemplo, a coletânea organizada por Shieber, 2004). Ainda assim, ele serve como uma boa introdução à história do teste e a algumas das questões centrais que ele coloca.

Em alguns momentos do livro, Mitchell ventila a ideia de que o que falta à IA simbólica e não-simbólica é levar mais a sério a importância do corpo na constituição da mentalidade, embora ela não desenvolva a ideia. Se voltamos às dificuldades enfrentadas no reconhecimento de objetos e as pensamos a partir das abordagens ecológica e enativa da percepção (Carvalho 2022), poderíamos dizer que falta a essas máquinas um corpo para explorar objetos de diferentes perspectivas e assimilar como a aparência desses objetos varia e é afetada pelos movimentos realizados. E não basta um corpo passivo, é preciso que seja um corpo ativo, que explore e possa então aprender a relação entre o que ele faz e o efeito disso na sua experiência. As contingências sensoriomotoras assim assimiladas embutem conhecimento causal do mundo e talvez seja incontornável para a obtenção de habilidades robustas de reconhecimento. Em um artigo mais recente, a autora explora algumas das lacunas da IA tradicional à luz da cognição corporificada (2021, p. 6-7). Ela explora a tese de que a familiaridade com o próprio corpo é crucial para a compreensão de conceitos abstratos (Lakoff & Johnson 2003) e a tese de que emoções são indispensáveis para decisões inteligentes (Damásio 2012). Os aspectos corporificados da mente, segundo a autora, começaram a ser explorados nas subáreas da IA corporificada e robótica do desenvolvimento, embora ainda sejam subáreas bem marginais na IA.

Embora Mitchell mencione o ataque de Searle ao programa forte da inteligência artificial, isto é, a ideia de que simular computacionalmente a mente é suficiente para gerar a mente — note o quão forte é essa afirmação, pois o mesmo claramente não se aplica à simulação computacional de um tornado —, ela não discute o célebre argumento do quarto chinês, o qual poderia reforçar o seu argumento central de que mesmo as IAs atuais, apesar de todo o relativo, como vimos, sucesso, não têm qualquer entendimento do que estão fazendo. Não apenas não têm conhecimento de segunda ordem sobre como fazem o que fazem, mas também, em um nível muito

mais básico, não têm intencionalidade, não são capazes de se direcionar e se referir ao que quer que seja. Pode ser que, para ter uma mente, precisemos não só de um corpo ativo, como sugeri acima, mas também um corpo vivo, no sentido em que a tradição fenomenológica o compreende, isto é, o corpo tal como o experienciamos na perspectiva da primeira pessoa, o corpo que toca, age e vê (Käufer & Chemero 2015, p. 99). Na verdade, pode ser que um corpo só possa ser genuinamente ativo, isto é, movido por propósitos e intencionalidade intrínsecas, se vivo, tanto no sentido biológico quanto no sentido fenomenológico. Essa é a aposta da visão enativa da mente, que entende que as categorias centrais para compreender a vida são também as categorias centrais para compreender a cognição e a inteligência. Isso não significa o fim da inteligência artificial, talvez ela ainda seja possível desde que seja possível a vida artificial (Di Paolo 2010). Não é o cérebro que precisa ser emulado para criar máquinas inteligentes, mas a própria vida, uma ideia que já tinha sido vislumbrada pela cibernética.

6. Medos e receios

Ao mesmo tempo em que a inteligência artificial desperta interesse e entusiamo, ela também desperta receio e medo. O medo mais comum, amplamente explorado na ficção científica, é o de que máquinas superinteligentes venham a nos dominar ou mesmo aniquilar. No entanto, a base para este temor é muito fraca. Não estamos minimamente próximos de produzir uma máquina que pensa, o que dirá de uma máquina superinteligente. Embora evangelistas da IA falem que o fenômeno da singularidade ocorrerá nas próximas décadas, essas afirmações são propagandas para alavancar investimentos para IA ou empolgação de momento. A singularidade se refere ao momento em que criamos uma máquina mais inteligente que nós, que, por sua vez, será capaz de criar uma máquina mais inteligente que ela mesma e assim sucessivamente, gerando em um curto espaço de tempo máquinas superinteligentes. Dado o que as máquinas atuais são capazes de fazer e a nossa compreensão da inteligência e do funcionamento do cérebro e da mente, é muito implausível que esse fenômeno ocorra nas próximas décadas.

Um receio mais plausível é o de que as máquinas venham a eliminar a maioria dos empregos atuais (Boden 2020, p. 216–218). Toda atividade que pode ser descrita por um processo algoritmo pode em princípio ser realizada por uma máquina. Os carros autônomos acabariam com milhares e milhares de empregos na área de transporte no mundo inteiro e, em um país como o Brasil, onde o transporte de mercadorias é sobretudo rodoviário e um número crescente de pessoas sobrevive como motorista de aplicativos, provocaria uma catástrofe social. Na área de serviços, recepcionistas, atendentes e agentes de telemarketing também poderiam ser nas próximas décadas

substituídos por máquinas. Alguns arriscam a dizer que até a aplicação da lei e o diagnóstico de doenças poderão ser feitos por máquinas, ameaçando empregos na área da saúde e do direito. Embora em outros episódios de revolução tecnológica a supressão de empregos tenha sido compensada pela criação de novos empregos relacionados à nova tecnologia, há o receio de que isso não se repetirá desta vez. A quantidade de empregos ameaçados pela IA é grande demais para que sejam compensados pelos empregos em torno dela. Além disso, os empregos que surgirão demandarão uma qualificação que muitos não possuem, especialmente nos países em desenvolvimento. Contudo, ainda é bastante incerto a extensão dos empregos que poderão ser confiavelmente substituídos por máquinas. Quanto aos carros autônomos, que talvez seja uma das subáreas da IA aplicada que mais concentra investimentos, Melanie Mitchell arrisca o prognóstico de que ainda estamos longe de carros plenamente autônomos. O mais plausível é que tenhamos, nas próximas décadas, carros parcialmente autônomos, requerendo motoristas humanos que estejam atentos e possam tomar o controle se algo sair errado, ou carros próximos da autonomia completa mas que estejam funcionando apenas em regiões planejadas para este fim, de modo a eliminar ou restringir a possibilidade de eventos inesperados ou sabotagem maliciosa. Isso se deve ao fato de que as redes neurais por trás desses carros não são robustas, como já vimos. Uma tarja preta numa placa de sinal que não impossibilita ainda um humano de identificá-la pode ser suficiente para causar um acidente com um carro autônomo. Assim, em países como o Brasil, não parece que o emprego de motorista esteja ameaçado no curto e médio prazo.

No início do livro, Melanie Mitchell ressalta que o maior receio de Hofstadter, autor de *Gödel, Escher, Bach: an Eternal Golden Braid* (1979), não era nenhum dos dois acima, mas o de que o avanço da IA poderia roubar a nossa unicidade e aniquilar com o nosso senso de humanidade. Na medida em que tarefas que eram feitas apenas por humanos passam a ser realizadas também por máquinas, de modo automático, nos veríamos como menos importantes. Quando algoritmos simples de força bruta mostram-se capazes de desafiar bons jogadores de xadrez, essa atividade que até então era vista como o pináculo da inteligência, passa a ser vista como algo banal que não confere mais tanto prestígio a quem a realiza bem. Não há nada de distintivamente humano em quem joga bem xadrez. De acordo com Melanie Mitchell, Hofstadter também ficou muito perturbado com o aplicativo EMI, produzido pelo músico David Cope, que é capaz de criar peças musicais no estilo de compositores como Bach e Mozart. As peças parecem tão genuínas que mesmo pessoas familiarizadas não conseguem distingui-las de peças originais. É como se até a criatividade, algo que nos distinguiria dos demais animais, tivesse sido exposta como banal, resultando de operações automáticas simples. Penso, no entanto, que o receio de Hofstadter está mal colocado. Por um lado, é muito precipitado inferir que jogamos xadrez ou criamos música da mesma maneira que essas máquinas. O mesmo efeito pode ter causas

muito distintas. Por outro, o senso de unicidade ou superioridade da humanidade já vai tarde, é algo de que deveríamos ter nos desvencilhado desde a descoberta da seleção natural. O que talvez Hofstadter tenha notado é a ameaça de que o avanço da IA nos faça ver a nós mesmos como também meros mecanismos. Não nos vermos mais como agentes é algo certamente a se temer. Mas essa não é uma dificuldade colocada particularmente pela IA ou o seu avanço, mas pela ciência moderna.

Por fim, para Melanie Mitchell, o seu maior receio em relação ao avanço da IA é que, devido ao deslumbramento diante dos seus feitos e à falta de compreensão dos seus limites e vulnerabilidades, acabemos presumindo que seus produtos e aplicações têm uma uma confiabilidade e robustez que eles não possuem e transfiramos para eles decisões importantes que ainda deveriam estar em nossas mãos. Não devemos temer máquinas superinteligentes, que não estão ainda ao alcance, mas máquinas burras e inseguras. É temerário que transfiramos para as máquinas decisões que elas não têm a inteligência para fazer. De todos os medos e receios em relação à IA, este não é apenas o mais fundamentado, ele é real, e aponta para uma realidade que já está presente em nossas vidas. E aqui reside uma questão filosófica incontornável, pois não há resposta tecnológica para a questão sobre como devemos incorporar a tecnologia em nossas vidas.

Referências

Adams, F.; Aizawa, K. 2001. The Bounds of Cognition. *Philosophical Psychology* **14**(1): 43–64.

Boden, M. A. 2020. *Inteligência Artificial: Uma brevíssima introdução*. São Paulo: Unesp.

Carvalho, E. M. 2022. Psicologia Ecológica: da percepção à cognição social. In: M. J. Alvez de Souza & M. M. de Lima Filho (ed.), *Escritos de Filosofia V: Linguagem e Cognição*, p.367–393. Porto Alegre: Editora Fi.

Damásio, A. R. 2012. *O erro de Descartes: emoção, razão e o cérebro humano*. Terceira ed. São Paulo: Companhia das Letras.

Damassino, N. 2020. The Questioning Turing Test. *Minds and Machines* **30**: 563–587.

Di Paolo, E. 2010. Robotics Inspired in the Organism. *Intellectica* **53**(1): 129–162.

Dreyfus, H. L. 1979. *What computers can't do : The limits of artificial intelligence*. New York: Harper & Row Publishers, Inc.

Hofstadter, D. 1979. *Escher, Bach: an Eternal Golden Braid*. New York: Basic Books.

Käufer, S.; Chemero, A. 2015. *Phenomenology: an Introduction*. Cambridge: Polity Press.

Lakoff, G.; Johnson, M. 2003. *Metaphors we live by*. Chicago: University of Chicago Press.

Mallory, F. 2020. In Defence of a Reciprocal Turing Test. *Minds and Machines* **30**: 659–680.

Mitchell, M. 2019. *Artificial intelligence: A guide for thinking humans*. New York: Ferrar, Straus and Giroux.

Mitchell, M. 2021. Why AI is harder than we think. In: GECCO '21: Proceedings of the Genetic and Evolutionary Computation Conference. Lille France: ACM. https://doi.org/10.1145/3449639.3465421. Access 09.11.2022.

Pearl, J. 2018. How a Pioneer of Machine Learning Became one of Its Sharpest Critics. https://www.theatlantic.com/technology/archive/2018/05/machine-learning-is-stuck-on-asking-why/560675/. Access: 15.05.2022.

Shieber, S. M. 2004. (ed.). *The Turing test: verbal behavior as the hallmark of intelligence*. Cambrigde, Mass.: MIT Press.

Turing, A. 1950. Computing Machinery and Intelligence. *Mind* **LIX**(236): 433–460.

Turing, A. 1951. Can Digital Computers Think? https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-5. Access: 03.08.2021.

Warwick, K.; Shah, H. 2016. Can machines think? A report on Turing test experiments at the Royal Society. *Journal of Experimental & Theoretical Artificial Intelligence* **28**(6): 989–1007.

Notas

[1]Veja https://www.newscientist.com/article/2176618-deepminds-ai-can-spot-eye-disease-just-as-well-as-top-doctors/? Acesso em 15/05/2022.

[2]Veja https://image-net.org/challenges/LSVRC/. Acesso em 15/05/2022.

[3]Veja https://www.mturk.com/. Acesso em 15/05/2022.

[4]Veja https://web.njit.edu/~ronkowit/eliza.html. Acessado em 05/01/2023.

[5]Veja https://www.botlibre.com/browse?id=857177. Acessado em 05/01/2023.

[6]Veja https://longbets.org/1/. Acesso em 15/05/2022.

Agradecimentos

Agradeço aos pareceristas anônimos e ao editor da revista pelos comentários ricos e precisos que ajudaram muito a melhorar esta resenha. Agradeço também aos estudantes do Seminário de Filosofia da Inteligência Artificial, ministrado em 2021/2 na UFRGS, pelas discussões animadas em torno do livro resenhado. Finalmente, agradeço ao CNPq pelo apoio financeiro, projeto n. 306795/2021-3.