ARTIFICIAL

MARIANO SIGMAN
SANTIAGO BILINKIS

ARTIFICIAL

A nova inteligência e a fronteira do humano

Tradução de
Ana Pinto Mendes

TEMAS E DEBATES

Sumário

Prólogo

Mariano Sigman

Num dia de primavera, em Madrid, o Emiliano Chamorro sugeriu-me que acrescentasse um capítulo a um livro que eu tinha escrito sobre conversação, um capítulo que falasse sobre como conversar com uma inteligência artificial. Foi aqui que tudo começou. A ideia viajou à velocidade de um raio, até chegar ao Miguel Aguilar e ao Roberto Montes, editores, mestres e amigos de um e outro lado do Atlântico. E regressou, quase no mesmo momento em que o Emi terminava a sua frase, com outra proposta: «Não seria melhor um livro novo?» E foi nesta vertiginosa sucessão que se decidiu que o livro, além disso, teria de ser escrito num ápice.

Sem parar um segundo para pensar, peguei no telefone, liguei ao Santiago, com quem me cruzo constantemente mas com quem nunca tinha colaborado, e propus-lhe escrever, a quatro mãos e com um prazo bastante apertado, um livro. Era como convidar para subir connosco o Evereste uma pessoa com quem nunca saímos sequer para caminhar. Como nesse dia nada era normal, constatou-se que o Santiago acabava de iniciar não sei quantos projetos que, por si só, já pareciam um abismo e, enquanto eu começava a pensar que este arroubo de loucura duraria o que costumam durar estes arroubos, ele disse que sim, que não sabia como, mas que íamos avançar. E avançámos.

Santiago Bilinkis

O telefonema do Mariano foi encontrar-me num momento de assoberbamento total: ao trabalho habitual como divulgador na rádio e à geração de conteúdos para o meu *podcast* e para as redes, somava-se o repentino interesse dos meios de comunicação por entender a revolução do *ChatGPT*. A inteligência artificial, um tema a que dedico uma grande parte do meu tempo nos últimos quinze anos e que só captava o interesse de alguns poucos *nerds*, estava de repente no centro da agenda pública. A grande meta desta etapa da minha vida, que é aproximar a tecnologia mais avançada da vida das pessoas, de uma forma que lhes pareça simples e estimulante, ganhava mais relevância do que nunca. A última coisa de que precisava nesse momento era de uma proposta tão extraordinária quanto irrecusável. E bastou o telefone tocar. A minha resposta ao Mariano foi instantânea. A ideia de trabalharmos juntos pareceu-me muito estimulante e combinar as nossas ideias num projeto conjunto era uma oportunidade maravilhosa. Mas isto não acaba aqui: poucas coisas me atraem tanto como uma meta impossível. Fazer um livro sobre inteligência artificial, escrevendo pela primeira vez a dois, com um coautor que vive noutro país, com cinco horas de diferença horária, e terminá-lo em poucas semanas… Impossível!

Onde é que tenho de assinar?

1

A génese da inteligência

Tragédia e esperança

Em maio de 1938, o almirante Sir Hugh Sinclair, dos serviços secretos britânicos, o MI6, comprou uma mansão construída no século XIX, conhecida como Bletchley Park. Tinha a localização ideal para criar um centro de operações: ficava a pouco mais de setenta quilómetros de Londres, perto de uma linha de comboio que passava pelas universidades de Oxford e Cambridge, e o esplendor arquitetónico do palácio ajudaria a camuflar as atividades secretas do governo durante a Segunda Guerra Mundial.

Pouco depois, os serviços secretos foram «à pesca» nas universidades mais importantes do Reino Unido, para recrutar uma formidável equipa de trinta e cinco físicos e matemáticos, que seriam liderados por Alan Turing e Dillwyn Knox. Assim, de forma abrupta e precipitada, se pôs em marcha esta sucursal secreta da Escola de Códigos e Criptografia do governo do Reino Unido. Logo que chegou aos esplêndidos jardins de Bletchley Park, o grupo de *nerds* ficou a saber qual seria a sua missão: nem mais, nem menos do que salvar o mundo. Os alemães utilizavam uma máquina chamada «Enigma», que encriptava as suas mensagens através de um sofisticado sistema de engrenagens baseado em três rotores que convertiam cada letra

numa outra. O objetivo de Turing e da sua equipa era decifrar este código. Era uma tarefa extremamente difícil, já que os nazis alteravam todos os dias a posição inicial dos rotores, dando origem a cento e cinquenta e nove triliões de combinações possíveis. Era preciso voltar a decifrar a posição todos os dias. Descodificar estas mensagens poderia fazer pender a balança da Segunda Guerra Mundial, porque permitiria aos Aliados aceder a informação sigilosa sobre os planos e ações inimigos.

Dado que grande parte dos homens jovens estava destinada ao campo de batalha, o governo britânico recrutou mais de seis mil mulheres para trabalhar em Bletchley Park. Falavam várias línguas e eram muito hábeis a jogar xadrez e a resolver palavras cruzadas. Entre elas estava Joan Clarke, que rapidamente se tornou uma das pessoas decisivas do projeto.

Turing, Clarke e a sua equipa trabalharam em contrarrelógio, pressionados pelo avanço do conflito bélico. Passadas algumas semanas, descobriram como decifrar as mensagens. Mas no momento em que compreenderam os cálculos e decisões necessários para decifrar o código da Enigma, perceberam também que era impossível resolvê-los a tempo. Encontraram a solução noutro dos recintos de Bletchley Park, onde o próprio Turing estava a desenvolver uma máquina de cálculo que recebeu o nome de «Bombe». Com a ajuda deste enorme dispositivo eletromecânico, criado em 1939 com base num velho projeto do matemático polaco Marian Rejewski, seria possível determinar o conteúdo das mensagens encriptadas pela máquina Enigma.

Os códigos nazis foram decifrados a tempo graças a uma assombrosa conjunção de fatores humanos e tecnológicos: por um lado, uma equipa privilegiada de mentes científicas que passaram, sem aviso prévio, de explorar universos abstratos numa

ardósia a salvar o mundo, e de pessoas que converteram o seu apreço pelos enigmas e as palavras cruzadas no principal recurso para decifrar o conteúdo das mensagens secretas do Reich. Por outro lado, consta que a vaidosa insistência dos nazis em usar repetidamente a fórmula «Heil Hitler» foi um erro colossal que simplificou a tarefa, já que é muito mais simples decifrar um código que contenha mensagens previsíveis que se repitam. E, por último, o aperfeiçoamento de dispositivos aparatosos capazes de executar a grande velocidade cálculos que os cérebros combinados destes cientistas não teriam realizado a tempo.

A Bombe não teria passado num teste de inteligência. Executava apenas um cálculo exigente e sofisticado para decifrar um enigma. Mas este esboço de pensamento humano depositado num dispositivo elétrico revelava já algumas características do que identificamos como inteligência. Fazia operações e tomava decisões que, até esse momento, eram realizadas apenas por pessoas «inteligentes». O programa idealizado por Turing para determinar a posição inicial dos rotores da Enigma foi uma versão muito rudimentar de uma inteligência artificial (IA).

Deste modo, nos dias de hoje, em que se entende geralmente a IA como algo oposto ao humano, será talvez bom recordar que o seu primeiro projeto embrionário foi motivado justamente pela urgência de salvar a humanidade do seu poder de autodestruição.

Que deus atrás de Deus começa a trama?

As operações em Bletchley Park foram encerradas no final da guerra. Os matemáticos e físicos que tinham sido recrutados nas melhores universidades regressaram a casa, tal como os soldados que regressam da sua missão de serviço. Contudo, ao

contrário destes últimos, os heróis e as heroínas de Bletchley Park não puderam dizer onde tinham estado nem o que tinham feito e carregaram o peso deste segredo a vida inteira. O final desta história épica surge assim velado por um manto triste e obscuro.

Como se uma descoberta fundamental para o desenlace da Segunda Guerra Mundial não tivesse sido suficiente para uma vida inteira, Turing prosseguiu as suas investigações sobre os temas mais intrincados e desafiantes da ciência. Num artigo seminal, intitulado «A base química da morfogénese», publicado sete anos após o final da guerra, revelou o mecanismo que dá origem aos padrões sofisticados da natureza, da forma das flores, ou de uma célula, às espirais dos caracóis. Seguindo esta premissa, a de mostrar que as coisas mais surpreendentes da vida emergem de regras simples, reatou a missão de entender a inteligência. E propôs-se imitá-la, retomando o projeto que tinha começado em Bletchley Park.

Passada a urgência bélica, Turing entendeu que o xadrez, um jogo que, historicamente, funcionou como metáfora do engenho humano, era um terreno adequado para estudar a inteligência num domínio restrito, mas relevante. «Deus move o jogador e, este, a peça / Que Deus atrás de Deus começa a trama / de pó e de tempo, sonho e agonias?», escreve Borges num poema em que repara na mesma analogia. O xadrez transformou-se, nesse momento, na cobaia da história da IA, foi o primeiro grande cenário para a sua exploração e desenvolvimento e, na atualidade, é o melhor terreno para observar o que acontece quando uma inteligência sobre-humana se estabelece em algum dos nossos domínios.

«Como se concebe um programa capaz de analisar uma posição no xadrez e tomar criteriosamente boas decisões?», questionou-se Turing. Para entender como funcionam os mecanismos da inteligência, baseou-se em si mesmo. Analisou

os seus raciocínios, tentando compreendê-los e extrapolá-los para uma máquina. Este foi o primeiro passo na procura de uma IA: emular e replicar a inteligência humana. Mais precisamente, a inteligência de Turing. Este exercício de pensar sobre o nosso próprio pensamento, conhecido como «metacognição», até esse momento só tinha sido alvo do interesse da psicologia, como uma busca para tornar explícito o processo através do qual raciocinamos.

O *Turochamp*, o primeiro programa de xadrez, nasceu em 1948 a partir de uma investigação desenvolvida por Turing e David Champernowne em Manchester. O programa funcionava como uma receita de culinária. Uma série de instruções sequenciadas definiam os passos para decidir um movimento, de uma forma tão bem especificada que poderia ser usado por qualquer pessoa, mesmo que nunca tivesse jogado xadrez.

Turing teve uma ideia semelhante à que tinha ocorrido a Leonardo da Vinci no século xv: o seu génio estava muito à frente do desenvolvimento tecnológico da época. Como o *Turochamp* estava acima das capacidades de *hardware* disponíveis, não podia contar com computadores capazes de executar o programa concebido. Turing reparou então que uma forma de resolver o problema era executar o programa no seu cérebro, seguindo uma após a outra as instruções indicadas pelo algoritmo. O *Turochamp* foi o primeiro programa de IA e foi executado num cérebro humano.

A sua capacidade de jogo revelou-se bastante medíocre. Além do mais, como programa, tinha grandes limitações: operava sobre um único domínio específico com regras muito claras (jogava xadrez, mas não sabia fazer mais nada, nem sequer um jogo muito mais simples) e dependia da clareza da linguagem formal usada pelo programador, da sua imaginação e do seu conhecimento do jogo.

O projeto nasceu além disso já com a morte anunciada: o facto de se conhecer em pormenor como o programa funcionava e de que forma resolvia cada decisão tornou-o menos atrativo, já que grande parte do nosso fascínio diante da inteligência humana radica, precisamente, no facto de não a compreendermos. Uma coisa que não seja enigmática, surpreendente e inexplicável não nos parece inteligente. E, desde logo pela sua estrutura, ao programa concebido por Turing faltavam estes elementos.

O *Turochamp* foi um marco histórico, mas os seus resultados nunca foram muito promissores. Sessenta e quatro anos depois, em 2012, no âmbito da celebração do centenário do nascimento de Turing, a Universidade de Manchester resgatou o algoritmo que ele tinha criado e colocou-o frente a frente com um dos melhores jogadores de todos os tempos: Garry Kasparov. O grande mestre russo arrasou o velho programa numa partida de dezasseis jogadas.

A fronteira do humano

Em 1950, Turing publicou um artigo académico em que apresentou, pela primeira vez, o enquadramento teórico do teste que conhecemos como «teste de Turing». As máquinas conseguem pensar? Ou, para tornar a pergunta mais precisa: conseguem pensar de uma forma indistinguível da de um ser humano?

Turing propôs um teste para responder a esta pergunta, baseado no jogo da imitação. Neste teste, um entrevistador, usando um terminal, alterna perguntas a dois *entes*: um é uma pessoa e o outro, um computador. Se o interrogador conseguir distinguir quem é a pessoa e quem é a máquina, o computador

não passa no teste de Turing. Se, pelo contrário, o computador conseguir confundi-lo, terá passado no teste.

O teste de Turing é engenhoso e estabelece um critério conciso para medir a inteligência das máquinas, mas tem vários problemas. Baseia-se numa ideia antropomórfica, já que presume que uma inteligência geral tem de se assemelhar a uma inteligência humana. Adicionalmente, ser capaz de se camuflar não é condição necessária nem suficiente para ser inteligente. Mesmo que fôssemos mais inteligentes do que um chimpanzé, não nos conseguiríamos fazer passar por um e, por conseguinte, não teríamos passado no teste de Turing. De igual modo, pode haver IA muito potentes que não consigam emular a inteligência humana e outras que a emulem sem que por isso sejam inteligentes.

Numa trágica ironia, Turing, que salvou o mundo de um horror inenarrável e dedicou a vida a estudar o raciocínio, foi condenado pela irracionalidade moral da sua época. Turing era homossexual, o que, naquele tempo, se considerava ser uma perturbação mental e um perigo para a sociedade. Em 1952, assaltaram-lhe a casa e, no âmbito da investigação, pressionaram-no até que confessasse que o ladrão era um amante. A vítima do roubo passou a ser acusada de «indecência grave» e, em vez de cumprir a pena na prisão — o que lhe teria custado o posto de investigador —, ele aceitou uma condenação com liberdade condicional e foi submetido a um processo de castração química, com uma série de injeções de estrogénios, para reduzir a sua libido sexual.

No dia 7 de junho de 1954, encontraram o corpo de Turing sem vida. Diz-se que, junto ao seu corpo, havia uma maçã meio comida, na qual, suspeita-se, teria antes injetado cianeto. Com apenas quarenta e dois anos, uma das mentes mais brilhantes da história da humanidade, um verdadeiro

promotor do mundo livre e democrático, morreu encurralado por este mesmo mundo livre ainda tomado pelo preconceito. Na génese da história da IA há uma profunda tragédia humana.

O equilíbrio nuclear

Durante a Segunda Guerra Mundial, a ciência invadiu o terreno da política, de um lado e do outro do Atlântico. Enquanto Turing, Clarke e um batalhão de mulheres linguistas, matemáticas e até peritas em palavras cruzadas impulsionavam a computação, a criptografia e a IA em Bletchley Park, do outro lado do oceano Albert Einstein escrevia a sua célebre carta a Franklin D. Roosevelt, então presidente dos Estados Unidos, que começava assim:

> Alguns trabalhos recentes de E. Fermi e L. Szilard, que me foram comunicados em forma manuscrita, levam-me a supor que o elemento urânio pode converter-se numa nova e importante fonte de energia no futuro imediato… Este novo fenómeno conduziria também à construção de bombas e é concebível, embora muito menos seguro, que se possam deste modo construir bombas extremamente potentes de um novo tipo.

Para explorar o potencial bélico deste tipo de armas, criou-se o Projeto Manhattan, cujo centro de operações era o Laboratório Nacional de Los Alamos, nos Estados Unidos, uma instalação que, à semelhança de Bletchley Park, na Inglaterra, funcionava em segredo e congregava os melhores cérebros da época. Ali, os mais destacados físicos na área da mecânica quântica e da física atómica, liderados por Robert Oppenheimer, trabalhavam no desenvolvimento de uma bomba nuclear. Estes

cientistas também desempenharam um papel decisivo no resultado da guerra.

Com o fim da guerra e o triunfo dos Aliados, as tecnologias destes dois projetos seguiram por caminhos muito díspares. Durante várias décadas, a IA transformou-se num campo de estudo periférico, curioso, mas sem transcendência, que interessava apenas a um grupo minoritário de entusiastas da tecnologia e da ficção científica. Pelo contrário, o armamento nuclear converteu-se no eixo fundamental do equilíbrio geopolítico e no fator decisivo da Guerra Fria.

Os Estados Unidos e a União Soviética desenvolveram os respetivos planos de armas nucleares e ambas as potências chegaram a contar com a mesma capacidade de reduzir a civilização a cinzas. Esta paridade serviu, em várias ocasiões, para travar uma escalada bélica de potenciais consequências terríveis. Era um equilíbrio nefasto e imbuído de incerteza, sem dúvida, mas, afinal, um equilíbrio. O matemático John von Neumann, que também foi um dos grandes pioneiros da computação e da teoria dos jogos, descreveu matematicamente este equilíbrio com uma fórmula que interpela a razão: 1 + 1 = 0.

O que é interessante é saber que não se chegou aqui por acaso. Na década de 1940, muitos dos especialistas em física nuclear dos Estados Unidos partilharam os seus saberes com a outra grande potência, num fabuloso *thriller* de espionagem. Alguns cientistas tornaram-se espiões da União Soviética porque apoiavam os ideais comunistas, mas outros fizeram-no seguindo o princípio da «paridade nuclear». Vislumbraram o futuro e compreenderam que, para evitar a aniquilação do planeta, era preciso garantir que nenhum país deteria o monopólio desse poder destrutivo. Deste modo, o precário equilíbrio entre as duas potências em confronto foi o resultado da decisão de um grupo muito pequeno de cientistas que consideravam que o conhecimento

devia estar nas mãos de ambos para, assim, alcançar esta situação de empate: a doutrina conhecida como «destruição mútua garantida». O norte-americano Ted Hall, licenciado em Harvard, e o cientista inglês Klaus Fuchs foram absolutos expoentes desta teoria da paridade nuclear. Convencidos de que era preciso igualar as condições de jogo para salvar a humanidade de um desastre, entraram em contacto com os soviéticos e mantiveram-nos informados sobre os progressos do Projeto Manhattan.

A visão deste grupo de cientistas, que entenderam que a distribuição da tecnologia nuclear determinaria o futuro do mundo e que eles tinham um papel decisivo e inevitável a desempenhar — por ação ou omissão — na configuração do mapa global, pode servir como guia para pensar acerca do avanço da investigação e do controlo sobre a IA no futuro próximo. Veremos neste livro que, atualmente, o poder de influência dos dois projetos tecnológicos se alterou: já não será o poder destrutivo da energia nuclear, mas antes o da IA, a ocupar o foco principal da cena política e económica.

Eliza, o primeiro vestígio humano numa máquina de silício

Enquanto o mundo estava pendente da tensão agreste entre duas potências nucleares, a IA continuava a ocupar um lugar bastante marginal na esfera das preocupações sociais. Naquele tempo, a IA não era nem de perto uma coutada de ricos e famosos. Nos seus redutos académicos, físicos, matemáticos e neurofisiólogos como Marvin Minsky, John Hopfield ou Warren McCulloch começaram a trabalhar na ideia das «redes neuronais», um conceito que permitiu vislumbrar de que modo a inteligência emerge a partir de um substrato que não é inteligente.

Nesta nova abordagem, já não se observava a inteligência humana para a escrever num programa, pretendendo-se pelo contrário ver se um cérebro digital e artificial era capaz de produzir comportamentos inteligentes. Superava-se assim uma limitação fundamental da formulação de Turing, já que a maioria das coisas que fazemos implica mecanismos que nos são inacessíveis a nós mesmos. A procura da inteligência através de redes neuronais não deixava de ser uma conceção antropocêntrica, mas implicava uma mudança cabal.

Aquilo que o cérebro humano tem de assombroso não radica na complexidade de um neurónio, mas antes nas camadas e formas como se organizam milhões deles. Um neurónio tem, essencialmente, uma tarefa muito simples: escuta os outros e, se estes gerarem um sinal suficientemente forte, então dispara e envia esse sinal a outros neurónios vizinhos. Forma-se assim um circuito muito simples entre unidades, capaz de codificar uma grande quantidade de padrões nas diferentes configurações de neurónios acesos e apagados. Veremos em breve de que modo as redes neuronais se converteram no motor da IA, mas antes queremos apresentar uma das primeiras celebridades desta disciplina. Chama-se *Eliza*. E é um programa.

Na mesma altura em que Minsky e Hopfield davam um empurrão a uma incipiente ciência da inteligência, o psicólogo norte-americano Carl Rogers andava preocupado com outra dimensão do pensamento humano: a loucura. Com o colega Abraham Maslow, mudaram a filosofia e a prática da psicoterapia, conferindo-lhe uma perspetiva mais humanista. Numa época em que a loucura era vista como uma doença pela qual as pessoas perdiam aquilo que nos torna humanos, Rogers propôs uma aproximação empática em psicoterapia e avisou que a loucura nos atravessa a todos. A loucura, pouco a pouco, começou a ser vista, e tratada, como uma de tantas facetas e expressões

no mundo diverso e diversificado do humano. Rogers fez da empatia o centro da relação terapêutica.

Estes dois universos, o da psicoterapia e da empatia, por um lado, e o das redes neuronais e da IA, por outro, podem parecer muito distantes, mas na realidade são bastante próximos. Ambos procuravam entender, por construção ou observação, a essência do humano. Não foi por acaso que estes dois mundos se encontraram num dos primeiros marcos da IA. Em 1966, inspirado pela noção de empatia desenvolvida por Rogers, meio século antes do *ChatGPT*, Joseph Weizenbaum, um professor de Informática do Instituto de Tecnologia do Massachusetts (MIT), programou *Eliza*, o primeiro *bot* conversacional da história. Tratou-se de uma tentativa de sintetizar um psicoterapeuta rogeriano, intrinsecamente empático.

Eliza conversava por escrito com o seu interlocutor e tentava dar a impressão de que compreendia os seus problemas e inquietações. Para conseguir que a conversa fosse o mais empática e o mais humana possível, o programa recorria a um algoritmo simples. Primeiro, reconhecia alguma palavra-chave e, com base nesta palavra, podia inferir o tema. Por exemplo, quando o interlocutor mencionava a palavra «pai» ou «mãe», *Eliza* reconhecia que estavam a falar sobre a família. Em seguida, fazia alguma pergunta ou comentário relacionado com este tema. Adicionalmente, tinha armazenadas frases feitas e também usava frases de continuidade, que eram úteis para simular que revelava interesse.

Vejamos em seguida um exemplo de como era um diálogo com *Eliza*:

Mulher jovem: Os homens são todos iguais.
Eliza: Em que sentido?
Mulher jovem: Estão sempre a aborrecer-nos por causa de alguma coisa.

Eliza: Ocorre-lhe algum exemplo concreto?
Mulher jovem: Bom, o meu namorado obrigou-me a vir aqui.
Eliza: O seu namorado obrigou-a a vir aqui?
Mulher jovem: Diz que passo a maior parte do tempo deprimida.
Eliza: Estou a ouvir que está deprimida.
Mulher jovem: É verdade. Sou infeliz.
Eliza: Acha que vir aqui a ajudaria a não ser infeliz?
Mulher jovem: Preciso de ajuda, disso tenho a certeza.
Eliza: O que significaria para si receber ajuda?
Mulher jovem: Se calhar podia aprender a dar-me bem com a minha mãe.
Eliza: Fale-me mais da sua família.
Mulher jovem: A minha mãe cuida de mim.
Eliza: Quem mais da tua família cuida de si?
Mulher jovem: O meu pai.
Eliza: O seu pai.
Mulher jovem: Nalgumas coisas é parecida com o meu pai.
Eliza: O que achas parecido?
Mulher jovem: Não é muito agressiva, mas acho que não quer que eu me aperceba disso.

Este programa, instalado num computador tão gigantesco quanto primitivo, baseado em meia dúzia de linhas de código de uma simplicidade estarrecedora, tornou-se uma estrela da conversação. Todos queriam falar com *Eliza*. Além da sua destreza circense e de ser a prova de que era possível uma máquina de silício conversar, demonstrava que a empatia e, com ela, uma das características essenciais da condição humana, é muito mais simples do que achamos. Um programa rudimentar, que

propõe unicamente que a pessoa continue a falar sobre o mesmo tema, gera a ilusão de ser empático.

Mas *Eliza*, tal como o *Turochamp*, não passaria num teste de inteligência rigoroso. Era incapaz de memorizar, não aprendia com as suas conversas, não entendia a ironia, tinha um sem-fim de temas sobre os quais não podia opinar e a sua conceção do que era compreender o interlocutor radicava unicamente em continuar a propor uma conversa sobre um mesmo tema. Também não passaria no teste de Turing, mas já poderia enganar o interlocutor durante algum tempo, simulando algo profundamente humano. E tinha uma coisa que o seu criador jamais teria imaginado: era apaixonante falar com ela.

Um cérebro profundo

A psicologia e a IA tiveram um bom ponto de encontro na empatia de *Eliza*, da mesma maneira que as redes neuronais de Hopfield tiveram um ponto de encontro com a neurociência. Como é que um programa informático aprende com base em estruturas neuronais? A resposta veio da principal teoria sobre a aprendizagem no cérebro, sintetizada na máxima que o canadiano Donald Hebb enunciou em 1949: *neurons that fire together wire together* («Os neurónios que disparam juntos, ligam-se»). Vemos aqui outro exemplo de um fenómeno emergente, como a formação de padrões estudada por Turing. Quando este mecanismo simples se aplica a uma grande rede, dá lugar a um vasto repertório de aprendizagens nas quais se cimenta a assombrosa complexidade da inteligência. É o sonho da engenharia e da ciência e, em certa medida, da arte: uma regra simples capaz de explicar e sintetizar as estruturas mais complexas e sofisticadas do universo.

Esta é a ideia essencial de uma rede neuronal. Uma malha o mais ampla possível, formada por diferentes camadas de neurónios idênticos. As possibilidades de combinação são tantas que permitem estabelecer circuitos capazes de codificar quase qualquer coisa. Cada padrão de ativação da rede, isto é, cada conjunto de neurónios que se ativam em simultâneo, estabelece uma representação «mental» de um objeto. Pode ser a representação de algo concreto, como um animal, ou de um ente abstrato. Estas estruturas, por sua vez, podem ser combinadas para formar representações mais complexas. Para dar um exemplo matemático: a ativação de um grupo de neurónios pode indicar se um número é par. A ativação de outro grupo de neurónios, se um número é maior do que cem. Estes dois circuitos podem combinar-se num outro novo para representar os números pares que, além disso, sejam maiores do que cem. Uma rede neuronal deste tipo estabelece uma relação unívoca entre os objetos e as suas representações em grupos específicos de neurónios. Os neurónios que se ativam quando a rede vê este objeto, segundo a regra de Hebb, ligam-se entre si. E neste entrelaçado em particular fica a recordação de um objeto que pode ser ativada e representada de maneira abstrata. Neste momento, enquanto lê estas páginas, estão a ser criadas novas ligações entre os neurónios do seu cérebro e estão a fortalecer-se outras que já existiam. Estas transformações na sua rede de neurónios constituem a forma segundo a qual se constrói a recordação desta leitura. O registo de informação numa rede neuronal artificial opera da mesma forma.

As redes neuronais artificiais organizam-se numa estrutura hierárquica de sucessivas camadas, outra ideia tomada ao cérebro humano. Na sua versão mais simples, incluem três camadas: uma que codifica a entrada, outra intermédia que a processa e representa de forma mais abstrata, e uma de saída para dar

uma resposta. Graças ao aumento da capacidade de computação do *hardware*, foi possível agregar uma cada vez maior quantidade de camadas intermédias, dando lugar a um novo tipo de rede neuronal, conhecida como aprendizagem profunda, ou pelo seu nome em inglês, *deep learning*.

Articulando uma enorme quantidade de camadas, este tipo de rede tornou-se sumamente potente e, pouco a pouco, começou a reduzir o grande fosso que o separava de um cérebro humano. Como todas as redes, estabelece representações (também chamadas atributos). As representações geradas numa camada servem de elementos para a fase seguinte, que consegue assim um nível de abstração maior. Esta característica torna-as muito potentes e começa a dotá-las de traços da inteligência humana, como o referido em relação à abstração.

O exemplo paradigmático, um dos mais estudados no nosso próprio cérebro e que serviu de campo de testes para a IA, é o da visão. No córtex visual há uma primeira camada que deteta os limiares onde a luminosidade ou a cor se alteram. Estes são os componentes básicos do sistema visual, os seus primeiros atributos. Em seguida, uma segunda camada pega nesta informação já processada e começa a combinar estes segmentos para codificar formas de geometria simples: um ângulo reto, um ângulo inclinado, um «T», um quadrado… Por sua vez, esta camada torna-se o *input* da seguinte, que a recombina para processar formas mais complexas, como uma cara, até alcançar codificações abstratas de objetos complexos (um gato, uma pessoa feliz, Pedro, um amanhecer de inverno). O resultado deste cálculo sequencial da rede é a identificação de todos os atributos que fazem com que um «gato» seja um gato. Isto permite que o cérebro o reconheça sem que importe se é adulto ou bebé, ou se está de perfil, deitado, desenhado por um pintor impressionista, curvado, adormecido ou a saltar… Todas estas imagens tão

diferentes correspondem ao mesmo conceito: têm em comum aquilo que define a essência do que é um gato. Este trabalho de abstração ou categorização é central para a inteligência e resolve-se de uma forma relativamente simples. Brutal no seu esforço computacional e nas centenas de milhões de neurónios necessários, mas simples na sua lógica e procedimento.

O aprendiz supera o mestre

As redes neuronais mudaram a forma de aprender das máquinas: já não são programadas com uma série de instruções escritas por um humano, sendo antes treinadas para que vão descobrindo os padrões de ligações neuronais que as tornam eficazes. Neste processo, surge um elemento que também está na essência da aprendizagem humana: a retroalimentação ou *feedback*. Regressemos ao exemplo que já vimos: uma rede neuronal tem de responder se uma imagem corresponde a um gato ou não. De início, as suas ligações são arbitrárias e, por conseguinte, o seu desempenho será quase ao acaso. Contudo, e é aqui que reside a chave, cada vez que receber a indicação de que acertou, o padrão de ligações que conduziu até esse acerto será reforçado, aumentando a probabilidade de essa resposta se repetir em situações similares. Pelo contrário, quando lhe for indicado que cometeu um erro, as ligações que conduziram a esse desacerto ficarão mais fracas, gerando o efeito oposto.

Deste modo, num processo laborioso, que, à velocidade de um computador, é possível em tempos razoáveis, a rede vai aprendendo a estrutura precisa de ligações que lhe permite resolver esta tarefa. Passado este treino, consegue responder com êxito a novas imagens que nunca viu. Neste momento, é

válido utilizar a metáfora de que a rede compreendeu o que é uma categoria. Para tal, terá formado algumas ligações específicas que correspondem aos atributos que deve utilizar para identificar essa categoria acertadamente. Este exemplo tão fácil de descrever, não de resolver, estende-se a quase qualquer problema que possamos associar à inteligência, mesmo os que sejam aparentemente mais sofisticados. Este mecanismo é uma versão simples do que se conhece como «aprendizagem por reforço» (reforçar os padrões que funcionam) e, por mais elementar que pareça, faz parte dos fundamentos da inteligência humana e artificial.

Surge aqui um achado surpreendente: ao aprender por sua conta com base neste processo, o aprendiz (a rede neuronal) consegue entender o problema melhor do que o mestre (o ser humano) que lhe apresentou estes casos ou, por outras palavras, adquirir uma capacidade sobre-humana para cumprir esta tarefa em particular. Fá-lo, identificando, para resolver o problema, atributos-chave que nós não levamos em linha de conta ou não conseguimos verbalizar. Surge assim outra surpresa: a maneira como uma rede neuronal resolve um problema pode tornar-se incompreensível para os humanos. A palavra «incompreensível» usa-se aqui num sentido literal. Do mesmo modo que um conjunto pode ser descrito por extensão (enumerando todos os elementos que o compõem) ou por compreensão (escrevendo uma regra que permita identificá-los), algo torna-se incompreensível quando não é possível expressar essa regra verbalmente. Deste modo, as máquinas podem fazer as coisas, mas não explicar-nos como as fazem e passam a ser um enigma para nós.

A receita para aprender através da retroalimentação que acabamos de apresentar é bastante simples. O problema é que a vida está repleta de situações em que ninguém pode dizer-nos

se o que fizemos está bem ou mal, mas em todo o caso é preciso aprender. Isto resolve-se, tanto no cérebro humano como nas redes neuronais artificiais, criando uma função de valor: uma representação abstrata de «quão bem se fez uma coisa». Por exemplo, no caso de um jogo, que é a sua versão mais simples, a função de valor de uma determinada posição representa a probabilidade de ganhar a partir desse ponto. Uma jogada é boa se nos conduzir a uma posição melhor, isto é, se aumentar essa probabilidade de ganhar. Por conseguinte, nesta versão de aprendizagem por reforço, o algoritmo procura descobrir as jogadas que melhoram a função de valor. Deste modo, a função de valor funciona como uma representação interna da retroalimentação. O algoritmo reforça ligações quando a função de valor aumenta e altera-as quando esta diminui. O programa realiza este processo de aprendizagem por si só, simplesmente a jogar. O que não é assim tão estranho. Muitos de nós aprendemos jogos sem ler as regras, observando apenas, experimentando e aprendendo com base no êxito ou no fracasso do que fizemos. A questão é que, fora dos jogos, pode ser muito difícil e arbitrário estabelecer esta função. Ainda assim, a chave é que o *input* humano para a rede neuronal é definir a função de valor, indicando-lhe o que deve maximizar. Em seguida, o algoritmo de aprendizagem por reforço resolve esta tarefa de forma muito eficaz. Depois de lhe dizermos o «quê», a IA encontra o «como».

Todos os jogos, o jogo

Estas redes neuronais, famosas por resolver todo o tipo de problemas da vida quotidiana, não são muito diferentes das concebidas por Hopfield há mais de quatro décadas. Yann Le Cun,