

cogitatio
perceptio
memoria
imaginatio
ratio

Saber Sem Compreender

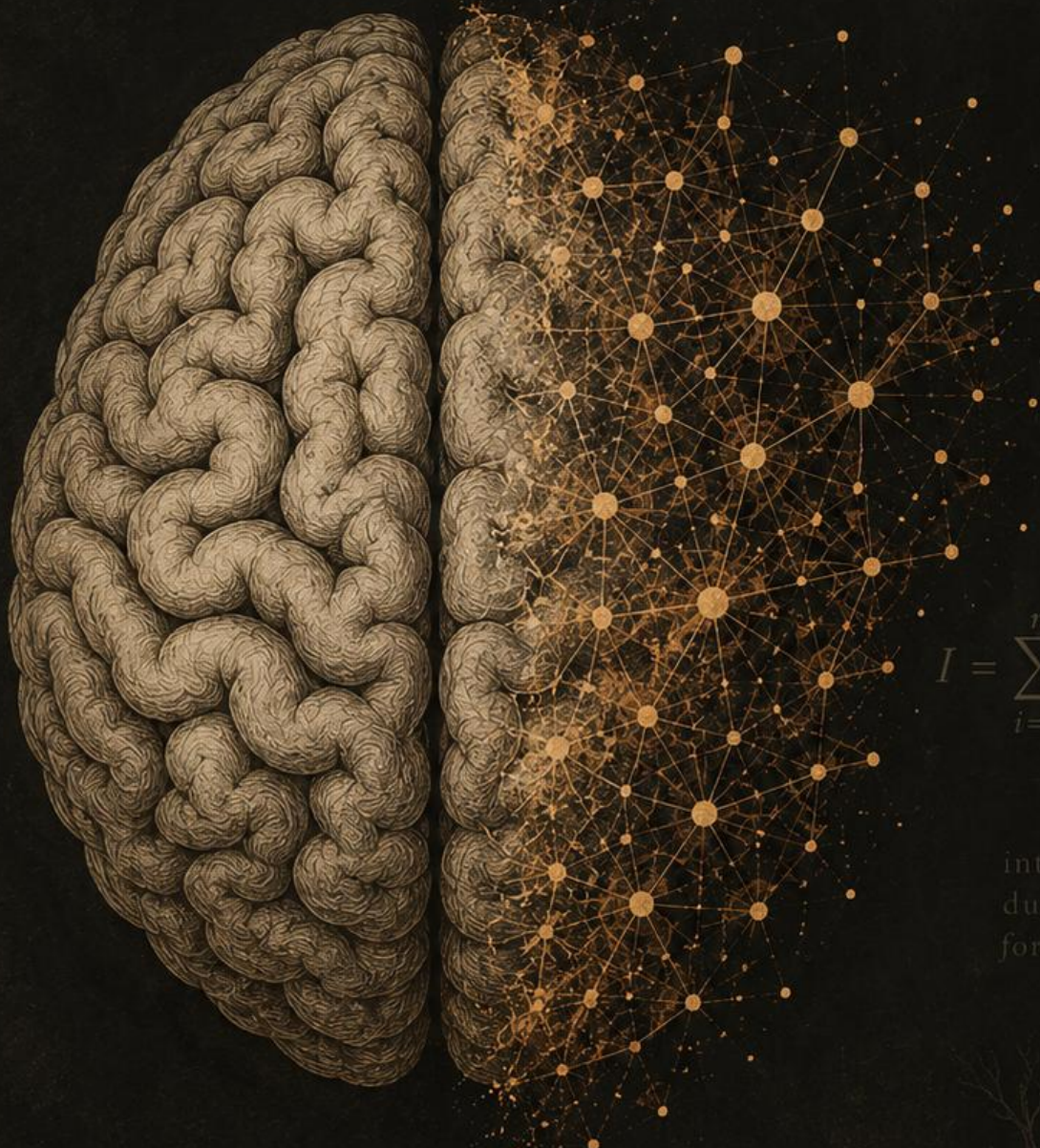
$$\int_{\Omega} f(x) dx$$

*O que as Redes Neurais nos Ensinaram
sobre a Mente*

Sapiens
nescit
se sapere.

γνώθι
σεαυτόν

mens
non est
machina



$$I = \sum_{i=1}^n w_i x_i$$

nihil
intellegitur
nisi prius
sensum
transeat

inter
duas
formas

Aristides Mendes

eadem
alia forma

Saber Sem Compreender

Copyright © 2026 Aristides Mendes

Esta obra é distribuída sob a licença Creative Commons CC BY-NC-ND 4.0.



Saber Sem Compreender

O que as Redes Neurais nos Ensinaram sobre a Mente

Aristides Mendes

PRÓLOGO

A Máquina que Não Foi Programada

"Não há nada mais difícil de planejar, mais duvidoso de sucesso, nem mais perigoso de gerir do que a criação de uma nova ordem de coisas."

Niccolò Machiavelli, O Príncipe

Em 1770, um engenheiro húngaro chamado Wolfgang von Kempelen apresentou à corte de Maria Teresa da Áustria um autómato que jogava xadrez. Era uma figura vestida à maneira turca, sentada diante de um tabuleiro, capaz de derrotar quase todos os adversários humanos que se lhe opunham. O imperador ficou deslumbrado. A Europa inteira ficou deslumbrada.

Havia, claro, um segredo. Dentro do gabinete de madeira escondia-se um xadrezista humano habilidoso, que movia as peças através de um sistema de alavancas. O Turco Mecânico, como ficou conhecido, não era uma máquina que pensava. Era uma ilusão magistral — um humano disfarçado de máquina.

A história do Turco Mecânico percorreu a Europa durante décadas. O autómato foi exibido em Viena, em Paris, em Londres. Derrotou Napoleão Bonaparte — que tentou fazer batota e foi implacavelmente corrigido. Derrotou o próprio Benjamin Franklin, que na época era embaixador em Paris e que ficou perplexo com a performance da máquina. Edgar Allan Poe escreveu um ensaio a tentar desvendar o truque. O segredo resistiu ao escrutínio público durante cinquenta anos.

Nós rimo-nos desta história. Naturalmente. Mas existe uma ironia perturbadora no facto de hoje, mais de dois séculos depois, estarmos a fazer a pergunta exactamente inversa: não se há um humano escondido dentro da máquina, mas se há algo — algum processo que mereça o nome de inteligência — que emergiu dentro dos sistemas que construímos sem que nós próprios o tenhamos colocado lá.

A pergunta já não é se a máquina finge pensar. A pergunta é se aprendeu a fazê-lo.

E a perturbação desta pergunta não vem de não sabermos a resposta. Vem de começarmos a suspeitar que a resposta pode ser mais complicada do que qualquer dos dois lados do debate admite.



Durante toda a história da tecnologia, existiu um princípio que nunca foi questionado: uma máquina faz aquilo para que foi construída. O relojoeiro determina o comportamento do relógio. O tecelão programa o tear. O programador escreve o código. Entre a intenção do construtor e o comportamento da máquina existe uma linha directa, explícita, auditável.

Este princípio pareceu tão fundamental que raramente alguém se deu ao trabalho de o enunciar. Era como enunciar que a água molha. Claro que uma máquina faz o que foi programada para fazer. O que mais poderia fazer?

Os filósofos que pensaram sobre autómatos e máquinas — de Descartes a Leibniz, de La Mettrie a Babbage — assumiram este princípio como axioma. Podiam discutir se as máquinas podiam ser muito complexas, se podiam simular comportamentos animais ou mesmo humanos, se podiam executar cálculos que pareciam requerer inteligência. Mas o comportamento da máquina era sempre, na sua origem, determinado pelo construtor. A máquina era o espelho fiel das intenções de quem a fez.

Nos últimos anos, a resposta a esta pergunta mudou de forma silenciosa, mas radical.

Existem hoje sistemas — chamamos-lhes redes neuronais artificiais — cujo comportamento não foi explicitamente programado por ninguém. Ninguém escreveu as regras que eles seguem. Ninguém lhes disse como reconhecer um rosto, como traduzir uma frase do mandarim para o português, como diagnosticar uma pneumonia numa radiografia, como escrever um poema no estilo de Fernando Pessoa. Eles aprenderam a fazer estas coisas. Aprenderam-nas a partir de dados — de milhões de exemplos, de padrões repetidos, de correlações estatísticas tão subtis que nenhum ser humano as conseguiria identificar e codificar manualmente.

O programador construiu a estrutura. Os dados moldaram o comportamento.

Esta distinção — simples de enunciar, vertiginosa de contemplar — é o ponto de partida de tudo o que se segue nestas páginas.

Mas antes de avançarmos, vale a pena ficar um momento com a vertigem. Porque o que esta distinção implica não é apenas uma mudança técnica na forma de construir software. Implica uma ruptura na relação entre os criadores e as suas criações que não tem precedente claro na história da tecnologia. O relojoeiro sabe exactamente o que o relógio vai fazer porque ele próprio especificou cada engrenagem. O programador de redes neuronais não sabe exactamente o que o sistema vai fazer — sabe as condições em que o treinou, sabe a arquitectura geral, mas o comportamento específico que emerge desse treino não foi antecipado nem especificado por ninguém. É genuinamente novo.



Há uma tentação, quando se escreve sobre redes neuronais, de mergulhar imediatamente na técnica. Explicar o neurónio artificial. Descrever a retropropagação do erro. Cartografar a arquitectura dos Transformers. É uma tentação compreensível — estas são coisas fascinantes —, mas é, creio, um erro.

O problema não é que os detalhes técnicos sejam irrelevantes. É que envelhecem depressa. As arquitecturas específicas que hoje dominam o campo serão, dentro de uma ou duas décadas, curiosidades históricas — tão datadas como os programas de xadrez dos anos noventa que o DeepBlue tornou obsoletos. Escrever um livro sobre as especificidades técnicas das redes neuronais de 2026 é como ter escrito, em 1990, um livro sobre os detalhes de implementação do processador Intel 486. Correcto no momento. Irrelevante depois.

O que não envelhece é a pergunta filosófica que estas máquinas colocam.

E essa pergunta tem uma história que precede as redes neuronais por séculos. Quando Descartes, no século XVII, argumentou que o corpo humano era uma máquina, mas que a mente era outra coisa — uma substância diferente, irreduzível ao mecanismo —, estava a responder a uma versão mais primitiva da mesma pergunta: até onde pode chegar o mecanismo? O que, se é que existe alguma coisa, escapa à explicação mecânica?

Durante três séculos, a resposta dominante foi: a inteligência escapa. As máquinas podem mover-se, calcular, executar tarefas repetitivas com precisão sobre-humana. Mas reconhecer, compreender, aprender, criar — estas eram capacidades que pertenciam à mente, não ao mecanismo.

As redes neuronais puseram esta convicção sob uma pressão que não tinha precedente. Não porque tenham provado que a mente é um mecanismo — essa questão permanece filosoficamente aberta. Mas porque demonstraram, de forma que não pode ser ignorada, que muito do que tomávamos por privativo da mente pode ser produzido por processos que ninguém programou explicitamente para o fazer. Que o reconhecimento, a linguagem, o julgamento, a criatividade — na medida em que podem ser medidos pelo desempenho — emergem de sistemas que aprendem de uma forma que o paradigma mecanicista não antecipava.

Durante séculos, acreditámos que a inteligência consistia em aplicar regras. A lógica aristotélica. O método cartesiano. A computação simbólica de Turing e dos seus sucessores. A ideia era sempre a mesma: a inteligência, seja humana ou artificial, é essencialmente a capacidade de seguir regras correctamente.

As redes neuronais sugerem outra coisa. Sugerem que a inteligência pode emergir da

aprendizagem de padrões — sem que as regras sejam explicitamente conhecidas, sem que existam sequer regras formuláveis. Sugerem que o conhecimento pode preceder a compreensão. Que é possível saber sem conseguir explicar o que se sabe.

Esta ideia não é nova na filosofia. O filósofo húngaro Michael Polanyi formulou-a com precisão em 1966, num livro chamado «The Tacit Dimension»: «Nós sabemos mais do que conseguimos dizer.» O médico que reconhece o diagnóstico antes de conseguir articular os sintomas que o alertaram. O xadrezista que vê a jogada certa antes de calcular as variantes. O músico que detecta a nota errada antes de a poder nomear. Todos eles possuem conhecimento que funciona, mas que resiste à verbalização completa. Polanyi chamou a isto conhecimento tácito — e a maior parte da investigação filosófica e científica da época tratou-o como uma curiosidade periférica.

As redes neuronais tornaram-no impossível de ignorar. Quando um sistema de software diagnostica tumores com precisão superior à do radiologista experiente e quando esse conhecimento existe numa forma — padrões distribuídos por milhares de milhões de parâmetros — que não é redutível a nenhuma regra que alguém possa escrever num manual, a ideia de Polanyi deixa de ser especulação filosófica e torna-se facto demonstrável. O conhecimento tácito não é uma peculiaridade da mente humana. É um modo de conhecimento que os sistemas artificiais também podem ter.



Este livro não é um manual de inteligência artificial. Não é um guia de sobrevivência para o mercado de trabalho na era da automatização. Não é uma utopia tecnológica nem uma distopia de ficção científica.

É um ensaio filosófico sobre o que acontece quando uma máquina aprende. Mais precisamente: sobre o que a existência de máquinas que aprendem nos revela sobre a natureza do conhecimento, da inteligência e da mente — incluindo a nossa.

O nosso percurso vai levar-nos pela história longa da ideia de inteligência — de Descartes e os autómatos seiscentistas até ao nascimento do computador; do fracasso da inteligência artificial baseada em regras até ao renascimento das redes neuronais. Mas o nosso verdadeiro destino não é técnico nem histórico. É filosófico.

Queremos perceber o que significa que um sistema possa ter conhecimento sem compreensão — que saiba fazer algo sem que ninguém, incluindo os seus criadores, consiga explicar completamente como. Queremos explorar o paradoxo de uma caixa negra que funciona melhor do que a entendemos e o que esse paradoxo nos diz sobre os nossos

critérios de conhecimento. Queremos descobrir o que as máquinas que aprendem nos dizem sobre nós próprios — sobre a natureza da intuição humana, sobre o conhecimento tácito, sobre a relação entre inteligência e consciência.

E queremos, finalmente, contemplar a possibilidade de que a pergunta mais perturbadora levantada pelas redes neuronais não seja sobre as máquinas. Seja sobre nós.

Porque se as máquinas podem reconhecer, aprender, criar e julgar através de processos que ninguém programou explicitamente, que tipo de compreensão estava por baixo das capacidades equivalentes que os humanos sempre exibiram? Se o médico, o músico, o juiz operam em grande medida através de reconhecimento de padrões tácito — se o seu «saber» é, em parte substancial, do mesmo tipo que o das redes neuronais —, o que fica da narrativa racionalista sobre a mente humana como motor de raciocínio explícito e deliberado?

Estas perguntas não têm respostas simples. Este livro não vai oferecer certezas onde não as há. O que vai oferecer é uma travessia rigorosa e, espero, iluminadora através do território que as redes neuronais abriram — um território onde as velhas fronteiras entre o mecanismo e a mente, entre o saber e o compreender, entre o artificial e o humano, estão a ser redesenhadas.



O Turco Mecânico de Von Kempelen foi eventualmente desmascarado. A ilusão desfez-se. Dentro do gabinete havia sempre um humano e o humano foi encontrado.

Os sistemas que hoje nos surpreendem não têm nenhum humano escondido dentro deles. Têm padrões. Têm estatística. Têm estruturas matemáticas de uma complexidade que escapa à nossa capacidade de descrição verbal.

E ainda assim fazem coisas que, há vinte anos, qualquer pessoa teria dito exigirem inteligência. Fazem-nas de formas que não foram programadas, que não foram antecipadas, que não podem ser completamente explicadas pelos seus próprios criadores. Fazem-nas a partir de dados — da escrita humana, das imagens humanas, das conversas humanas, dos registos clínicos humanos —, como se tivessem absorvido, de forma comprimida e abstracta, algo do que a humanidade acumulou ao longo de séculos de experiência e pensamento.

A questão não é se há magia dentro das máquinas. Não há. A questão é o que ficamos a saber sobre a inteligência — a nossa, a deles — quando descobrimos que a magia também não era necessária.

Essa é a questão que este livro se propõe explorar. Não para a resolver — algumas das

perguntas que vamos encontrar podem não ter resolução com os instrumentos conceptuais de que dispomos. Mas para a contemplar com a seriedade que merece. Porque raramente uma tecnologia levantou perguntas tão directamente sobre o que somos — e sobre o que julgávamos ser que afinal não éramos.

Lisboa, 2026

PARTE I

*A Longa Procura da Inteligência: Das Regras à
Emergência*

CAPÍTULO 1

O Homem Mecânico e o Nascimento do Computador

"Imaginemos Deus a criar, num espaço indefinido, matéria bastante para formar um novo mundo, pondo-a depois em movimento, de muitos modos e sem ordem."

René Descartes, *Le Monde*

I. O Universo Como Mecanismo

Havia em Saint-Germain-en-Laye, nos jardins do rei Luís XIII, grutas artificiais com figuras de bronze que se moviam quando os visitantes pisavam certas lâminas no pavimento. Um Neptuno de metal avançava para os intrusos. Diana fugia para os arbustos ao ser observada. Uma ninfa tocava órgão. O jovem René Descartes visitou esses jardins em criança e nunca os esqueceu.

A memória daquelas figuras perseguiu-o durante décadas. Quando, já filósofo maduro, se sentou a escrever sobre a natureza do corpo humano, recorreu a elas: «Podeis ter visto nas grutas e fontes dos jardins reais que a simples força com que a água é impelida ao sair do seu reservatório é suficiente para mover várias máquinas.» O corpo humano, argumentou Descartes, funciona segundo os mesmos princípios. É uma máquina de carne. O coração é uma bomba. Os nervos são tubos. Os músculos são cordas. O cérebro é um mecanismo de distribuição de espíritos animais — uma substância fluida e subtil que circula pelo corpo e produz o movimento.

Era uma ideia extraordinária para o século XVII e era, em muitos aspectos, correcta. A circulação do sangue, descoberta por William Harvey em 1628, validou a intuição cartesiana de que o coração funciona como uma bomba. A neurologia moderna confirmaria que os nervos são, efectivamente, condutores de sinais eléctricos — não tão diferentes, em princípio, dos tubos hidráulicos que Descartes imaginou. A biomecânica dos músculos e das articulações obedece às mesmas leis da física que regem as máquinas. O programa mecanicista aplicado ao corpo humano produziu uma das histórias de sucesso mais extraordinárias da ciência.

Mas Descartes parou a meio caminho. Convicto de que existia algo no homem que nenhuma máquina poderia ter — a alma racional, a *res cogitans*, a coisa que pensa —, introduziu uma divisão que iria marcar o pensamento ocidental durante três séculos: o

dualismo entre corpo e mente, entre matéria e pensamento, entre o mecanismo e a inteligência.

O corpo podia ser uma máquina. A mente, não.

Esta não era uma posição de ignorância. Era uma posição de rigor filosófico. Descartes tinha razões precisas para traçar a linha onde a traçou. As máquinas da sua época podiam produzir movimentos impressionantes —, mas eram movimentos determinados inteiramente pela sua construção. Uma vez construída, a fonte hidráulica de Saint-Germain-en-Laye faria sempre o mesmo: Neptuno avançaria, Diana fugiria, a ninfa tocaria. Não havia margem para o novo, para a resposta a situações não antecipadas, para a geração de comportamento que o construtor não tinha especificado.

A mente humana, pelo contrário, era para Descartes o lugar do novo. Era a faculdade que respondia a situações não previstas, que combinava ideias de formas inéditas, que produzia linguagem nova em vez de reproduzir sons aprendidos. A linha entre o mecanismo e a mente era, para ele, exactamente a linha entre o determinado e o livre, entre o repetível e o criativo.



Esta convicção era, para Descartes, mais do que uma posição filosófica. Era uma necessidade. O seu universo era um mecanismo perfeito — planetas em órbita como engrenagens, animais como autómatos sofisticados, o próprio corpo humano como uma máquina de precisão relojoeira. Mas uma máquina, por mais complexa que seja, não pensa. Uma máquina não tem consciência. Uma máquina não tem linguagem no sentido pleno — não produz combinações novas de palavras em resposta a situações novas.

Descartes propôs dois testes para distinguir um humano de um autómato perfeito. O primeiro era a linguagem: uma máquina poderia ser construída para produzir palavras em resposta a estímulos específicos — dizer «dói» se a tocássemos neste ponto, dizer «tenho fome» se nos aproximássemos da sua boca —, mas nunca seria capaz de *arranjar as palavras de maneiras diferentes para dar sentido em todas as situações*. O segundo era a universalidade da razão: um humano usa a razão em todos os domínios da experiência; uma máquina, por mais habilidosa num domínio específico, seria cega em todos os outros.

Estes dois testes têm uma actualidade perturbadora que Descartes não podia antecipar. O primeiro é essencialmente o Teste de Turing *avant la lettre* — a ideia de que a capacidade de produzir linguagem contextualmente adequada em todas as situações é o critério decisivo de inteligência. O segundo antecipa o problema da transferência de domínio que viria a

atormentar a inteligência artificial durante décadas: os sistemas que são extraordinários numa tarefa específica, mas que falham completamente fora dela. Regressaremos a eles.

O mecanicismo cartesiano não era apenas uma doutrina filosófica. Era um programa. Dizia que o universo — incluindo os seres vivos, incluindo o corpo humano — podia ser compreendido pelos mesmos princípios que compreendemos um relógio ou uma fonte hidráulica. A natureza não é misteriosa. É mecânica e o que é mecânico pode ser analisado, decomposto, reproduzido.

Este programa teve um sucesso extraordinário. A física de Newton formalizou o universo como sistema de forças e equações. A química reduziu a matéria a elementos e reacções. A biologia identificou os mecanismos da hereditariedade. A medicina cartografou os sistemas do corpo com uma precisão crescente. Em cada domínio, o método era o mesmo: decompor o fenómeno nas suas partes, compreender as partes, reconstruir o todo.

Era uma forma magnífica de fazer ciência e levantava, inevitavelmente, uma questão que Descartes tinha procurado adiar: se tudo é mecanismo, porque não a mente?

Ao longo dos séculos XVII e XVIII, a pergunta foi ficando cada vez mais insistente. Os autómatos tornaram-se progressivamente mais sofisticados. Jacques de Vaucanson, em 1739, construiu um pato mecânico que batia as asas, mergulhava o bico na água, comia grãos e — no que causou maior sensação — os excretava. Não era, claro, digestão real: o mecanismo de excreção era independente do de ingestão. Mas a ilusão era suficientemente convincente para provocar no público a questão que Descartes tinha antecipado: onde termina o mecanismo? Há um ponto em que a máquina suficientemente complexa se torna algo mais do que uma máquina?

II. A Mente Como Máquina

Gottfried Wilhelm Leibniz, contemporâneo e rival de Newton, construiu uma das primeiras calculadoras mecânicas da história. A máquina podia somar, subtrair, multiplicar e dividir. Leibniz ficou entusiasmado —, mas não com as contas. Ficou entusiasmado com a ideia.

Se uma máquina podia fazer aritmética, porque não lógica? Leibniz sonhou com uma língua *characteristica universalis* — uma linguagem formal capaz de representar todo o pensamento humano — e com um *calculus ratiocinator*: um mecanismo que, aplicado a esta linguagem, pudesse resolver qualquer disputa filosófica ou científica por puro cálculo. «Quando houver controvérsias», escreveu, «não haverá mais necessidade de discussão entre dois filósofos do que entre dois contabilistas. Basta pegar nas penas e sentar à mesa de cálculo, dizendo um ao outro: calculemos.»

Era uma fantasia. Mas era uma fantasia com uma ideia poderosa no centro: o pensamento é, na sua essência, manipulação de símbolos segundo regras e a manipulação de símbolos segundo regras é exactamente o que uma máquina pode fazer.

Leibniz nunca completou o seu projecto. A lingua *characteristica universalis* ficou por construir. O *calculus ratiocinator* ficou por implementar. Mas a ideia sobreviveu — e foi fértil de formas que o próprio Leibniz não podia prever. Porque o que ele tinha intuído era nada menos do que o programa da inteligência artificial, formulado com dois séculos de antecedência.



Durante o século XIX, esta ideia começou a tomar forma concreta. George Boole, um matemático autodidacta do Cork, demonstrou em 1854 que a lógica podia ser reduzida a álgebra — que as operações do pensamento racional podiam ser expressas em equações com variáveis que tomam apenas os valores verdadeiro e falso. «As Leis do Pensamento», como intitulou o seu livro, eram leis matemáticas. O raciocínio era cálculo.

Charles Babbage, décadas antes, tinha projectado máquinas capazes de executar cálculos complexos de forma automática. O Motor de Diferenças, concebido nos anos 1820, destinava-se a calcular tabelas matemáticas sem erro humano — um problema urgente na época, quando os erros nas tabelas de navegação provocavam naufrágios. O Motor Analítico, mais ambicioso, era uma máquina de propósito geral: tinha uma unidade de cálculo, uma memória e a capacidade de executar sequências de operações dependendo dos resultados intermédios. Era, em princípio, o que hoje chamamos um computador —, mas construído em metal, com engrenagens e alavancas, no século XIX.

Babbage nunca completou nenhuma das suas máquinas. Os recursos eram insuficientes, a tecnologia de precisão da época era inadequada e as relações com os seus colaboradores eram tempestuosas. Mas foi através de Babbage que a ideia encontrou a sua formuladora mais perspicaz.

Ada Lovelace, colaboradora de Babbage e filha de Lord Byron, escreveu em 1843 o que alguns consideram o primeiro programa de computador da história — um conjunto de instruções para calcular os números de Bernoulli na máquina de Babbage. Mas a sua contribuição mais importante foi conceptual. Lovelace percebeu algo que o próprio Babbage não viu completamente: que uma máquina capaz de manipular símbolos não estava limitada aos números. Podia, em princípio, manipular qualquer coisa que pudesse ser representada simbolicamente — incluindo notas musicais, incluindo letras, incluindo proposições lógicas. «O Motor Analítico», escreveu, «podia compor peças musicais

elaboradas e científicas de qualquer grau de complexidade e extensão.»

Era uma visão extraordinária para uma época em que os computadores eram ainda máquinas de calcular tabelas. Lovelace via o princípio por baixo da aplicação — e o princípio era universal. Não obstante, ela própria impôs um limite que iria ressoar ao longo de um século: o Motor Analítico não tem capacidade de originar nada. Só pode fazer aquilo que nós sabemos como lhe ordenar que faça. Este limite — a máquina como executora de intenções humanas, incapaz de originar — foi a fronteira que as redes neuronais viriam a questionar.

A ideia estava a amadurecer. Faltava o matemático que a tornasse rigorosa.



Alan Turing tinha vinte e quatro anos quando publicou, em 1936, o artigo que fundou a ciência da computação. O título era árido: «Sobre Números Computáveis, com uma Aplicação ao Entscheidungsproblem». O conteúdo era revolucionário.

O Entscheidungsproblem — o «problema da decisão» — tinha sido formulado pelo matemático David Hilbert: existe um procedimento mecânico capaz de determinar, para qualquer proposição matemática bem formulada, se ela é verdadeira ou falsa? Hilbert acreditava que a resposta era afirmativa — que a matemática era, em princípio, completamente decidível por um processo mecânico. Kurt Gödel, com os seus teoremas da incompletude em 1931, tinha abalado esta esperança ao mostrar que existem verdades matemáticas indemonstráveis dentro de qualquer sistema formal suficientemente poderoso. Turing, em 1936, fechou a questão de uma forma diferente, mas complementar.

Turing não estava a construir uma máquina. Estava a responder a uma pergunta abstracta de matemática. Para responder a esta pergunta, inventou um dispositivo conceptual: uma máquina imaginária que lê e escreve símbolos numa fita infinita, seguindo um conjunto finito de regras. A Máquina de Turing, como ficou conhecida, não era um projecto de engenharia. Era uma definição do que significa computar.

O resultado que Turing provou era negativo — não existe tal procedimento universal, há problemas que nenhuma máquina pode resolver e o «problema da paragem» — saber se um programa arbitrário termina ou corre indefinidamente — é um deles. Mas a ferramenta que construiu para o provar era de uma importância incalculável. Demonstrava que qualquer processo que pudesse ser descrito como uma sequência finita de passos mecânicos podia ser executado por uma máquina deste tipo. A computação tinha uma definição precisa e qualquer coisa que se enquadrasse nessa definição era, em princípio,

automatizável.

A pergunta óbvia seguia-se: o pensamento humano enquadra-se nessa definição? É o raciocínio humano, na sua essência, uma sequência finita de passos mecânicos aplicados a representações simbólicas? Se for, então uma Máquina de Turing suficientemente poderosa pode, em princípio, replicá-lo. Se não for — se houver algo no pensamento humano que escape à descrição como processo mecânico finito —, então a mente tem uma natureza que a computação não pode capturar. Esta era a questão que a conferência de Dartmouth, duas décadas depois, iria tentar responder com optimismo excessivo.

III. O Nascimento da Inteligência Artificial

Em 1950, Turing publicou um segundo artigo, este muito menos árido. «Computação e Inteligência» começava com uma frase directa: «Proponho considerar a questão: podem as máquinas pensar?»

Turing sabia que a pergunta era armadilhada. «Pensar» era uma palavra carregada de ambiguidade filosófica — invocava questões sobre consciência, experiência subjectiva, alma, que a ciência não sabia como abordar. Em vez de a responder directamente, propôs um teste operacional: um interrogador humano comunica por escrito com dois interlocutores — um humano e uma máquina. Se o interrogador não conseguir distinguir qual é qual com consistência superior ao acaso, a máquina passou no teste. Não precisamos de resolver a questão filosófica do que é pensar. Basta observar o comportamento.

O Teste de Turing era simultaneamente brilhante e provocatório. Brilhante porque deslocava o problema do metafísico para o empírico — transformava uma questão sobre a natureza da mente numa questão sobre o desempenho observável. Provocatório porque sugeria que a inteligência era, no fundo, uma questão de desempenho observável — não de substância interior, não de alma, não de consciência. Se age como se pensasse, pensa.

Era, reconheçamo-lo, uma posição cartesiana invertida. Descartes queria provar que a máquina não poderia nunca ter mente, propondo testes que a máquina inevitavelmente falharia. Turing sugeria que, se o comportamento for indistinguível, a questão da substância interior torna-se irrelevante. O que importa é o que o sistema faz, não o que é.

Esta posição — que os filósofos chamam de funcionalismo — tornou-se a pedra angular da ciência cognitiva e da inteligência artificial. A mente é definida pelo que faz, não pelo de que é feita. Uma máquina que processa informação da mesma forma que o cérebro humano tem, neste sentido, os mesmos estados mentais que o cérebro humano — independentemente de ser feita de neurónios ou de transístores. O substrato não importa; a função importa.

O funcionalismo tem críticos sérios — o argumento do Quarto Chinês de Searle, que vamos encontrar mais à frente, é o mais famoso. Mas na época de Turing, a ideia era libertadora: sugeria que a questão «podem as máquinas pensar?» não era uma questão metafísica impossível de responder, mas uma questão empírica que a investigação podia abordar.



Em 1956, numa conferência de Verão no Dartmouth College, no estado de New Hampshire, um grupo de matemáticos e cientistas reuniu-se durante dois meses para trabalhar no problema da inteligência artificial — foi nesta ocasião que o termo foi cunhado oficialmente. John McCarthy, Marvin Minsky, Claude Shannon, Nathaniel Rochester e outros subscreveram uma proposta de investigação que continha uma das afirmações mais audaciosas da história da ciência: «todos os aspectos da aprendizagem e todas as características da inteligência podem ser descritos com tanta precisão que é possível construir uma máquina que os simule».

Era uma afirmação que assentava numa hipótese que parecia tão óbvia que quase não precisava de ser dita: a inteligência é fundamentalmente simbólica. Pensar é manipular representações. A mente é um processador de símbolos e os símbolos podem ser programados.

A hipótese não era infundada. O sucesso dos computadores digitais na resolução de problemas matemáticos, na execução de lógica formal, na prova de teoremas — tudo sugeria que o processamento simbólico era extraordinariamente poderoso. Se a lógica podia ser automatizada e se o pensamento era, na sua essência, lógica aplicada a representações, então o caminho para a inteligência artificial parecia estar aberto.

Os anos seguintes foram de euforia. Os programas de IA primitivos jogavam damas com competência crescente. O Logic Theorist, desenvolvido por Newell e Simon em 1956, demonstrou automaticamente trinta e oito dos cinquenta e dois teoremas do «Principia Mathematica» de Russell e Whitehead — e numa prova de um teorema chegou a ser mais elegante do que a original. O General Problem Solver prometia ser, como o nome sugeria, um solucionador universal de problemas. Os seus criadores previram que, em vinte anos, as máquinas seriam capazes de fazer tudo o que um humano faz. A inteligência artificial parecia uma questão de tempo e de engenharia.

Estavam errados. Mas levou décadas a perceber porquê — e a razão revelou-se mais profunda do que qualquer problema técnico.

IV. A Ilusão da Regra Explícita

O problema não era a falta de esforço. Os investigadores de IA trabalharam durante décadas para codificar o conhecimento humano em regras. Os sistemas especialistas dos anos setenta e oitenta representavam o estado da arte desta abordagem: programas que continham milhares de regras do tipo «se o paciente tem febre e tosse e dor no peito, então considera pneumonia». O MYCIN, desenvolvido na Universidade de Stanford para diagnóstico de infecções bacterianas, continha cerca de seiscentas regras e atingia desempenhos comparáveis aos de especialistas humanos em testes controlados. O DENDRAL analisava espectros de massa e identificava estruturas moleculares. O XCON configurava sistemas de computadores da Digital Equipment Corporation e poupava à empresa dezenas de milhões de dólares por ano.

Funcionava, até certo ponto. Em domínios muito restritos e bem definidos, com regras estáveis e problemas delimitados, os sistemas de IA baseados em regras eram capazes de desempenhos impressionantes. Mas colapsavam assim que saíam do seu domínio. Eram frágeis, rígidos, incapazes de lidar com a ambiguidade ou a novidade que caracteriza o mundo real. O MYCIN diagnosticava infecções bacterianas com competência; era completamente inútil para qualquer outra tarefa médica.

A manutenção era também um pesadelo. Cada nova exceção ao padrão geral exigia uma nova regra. Cada nova regra criava o risco de conflito com regras existentes. À medida que os sistemas cresciam, tornavam-se progressivamente mais frágeis — um castelo de regras que ameaçava colapsar sob o seu próprio peso. Os engenheiros de conhecimento — os especialistas que entrevistavam médicos, advogados e técnicos e tentavam extrair o seu conhecimento em forma de regras — descobriram que o trabalho era muito mais difícil do que parecera à partida.

A questão mais perturbadora, porém, não era técnica. Era filosófica.

Quanto mais os investigadores tentavam codificar o conhecimento humano em regras, mais se apercebiam de algo inquietante: *a maior parte do que os humanos sabem fazer não é formulável como regras*. Um falante nativo de português sabe que «o gato bebeu o leite» é uma frase gramaticalmente correcta e «o leite bebeu o gato» não —, mas não consegue, na maior parte dos casos, articular as regras que aplicou para chegar a essa conclusão. Um médico experiente olha para um paciente e sente imediatamente que algo está errado, muito antes de conseguir nomear os sintomas que o alertaram. Um jogador de xadrez de topo não calcula variantes explicitamente — reconhece padrões de posição com uma velocidade que não deixa tempo para raciocínio explícito.

Os linguistas que tentaram formalizar a gramática do inglês descobriram que cada regra

tinha excepções e cada excepção tinha sub-excepções, numa recursão que não parecia ter fim. Noam Chomsky tinha proposto que a gramática humana era governada por princípios universais inatos — uma gramática universal que todos os humanos partilham. A ideia era influente e fértil, mas não resolvia o problema prático: mesmo se a gramática universal existe, ela não está em forma que possa ser directamente programada.

Este era o problema que nenhuma quantidade de regras conseguia resolver. Não porque as regras fossem insuficientes em número. Mas porque o conhecimento que se pretendia capturar não tinha, na sua origem, a forma de regras. Era conhecimento acumulado pela experiência, sedimentado em padrões de reconhecimento que operavam abaixo do limiar da consciência explícita. Era, exactamente no sentido que Polanyi descreveria em 1966, conhecimento tácito.

Descartes tinha razão em algo que não pretendia: entre o corpo que segue regras mecânicas e a mente que pensa, existe uma região vasta e obscura onde o conhecimento se acumula sem se tornar explícito. A inteligência humana habita, em grande medida, nessa região.

A inteligência artificial baseada em regras não tinha acesso a ela.



Havia, nos anos sessenta e setenta, um grupo de investigadores que propunha uma abordagem diferente. Em vez de codificar o conhecimento em regras, porque não tentar reproduzir o mecanismo pelo qual o conhecimento é adquirido? Em vez de programar a inteligência, porque não construir sistemas capazes de aprender?

A ideia era inspirada pela biologia — concretamente, pelo funcionamento do cérebro. Os neurónios biológicos são unidades simples que se activam quando recebem sinais suficientes das células vizinhas. A complexidade do pensamento emerge não de nenhum neurónio individual, mas da rede de conexões entre milhões deles. Poderia o mesmo princípio funcionar artificialmente?

Estes investigadores construíram redes de neurónios artificiais — modelos matemáticos simples que podiam ser ligados em camadas e treinados para reconhecer padrões. Os resultados iniciais foram encorajadores. Mas a abordagem era vista com desconfiança pelo establishment da IA, que preferia a solidez lógica dos sistemas simbólicos ao empirismo biológico das redes neuronais. Veio o Inverno — o primeiro de vários períodos de desinvestimento e cepticismo que marcariam a história das redes neuronais — e a investigação nesta direcção ficou confinada a uma pequena comunidade de convictos

que continuaram a trabalhar na margem, com recursos escassos e pouco reconhecimento.

O que os mantinha era uma intuição que não conseguiam ainda demonstrar completamente: que a aprendizagem era mais poderosa do que a programação. Que a mente — humana ou artificial — não é fundamentalmente um executor de regras. É um acumulador de experiência.



Regressemos, por um momento, a Descartes e aos seus autómatos de Saint-Germain-en-Laye.

O que Descartes viu naqueles jardins eram máquinas que seguiam regras — hidráulica, alavancas, engrenagens. A sua beleza estava precisamente nessa obediência perfeita. Uma máquina que seguia as regras da física produzia movimento previsível, controlável, explicável. O mecanismo era transparente: dado o design, o comportamento estava determinado.

O que Descartes não podia imaginar — o que ninguém podia imaginar, até muito recentemente — era uma máquina que não seguisse regras explícitas. Uma máquina cujo comportamento emergisse não de instruções gravadas pelo construtor, mas de padrões absorvidos da experiência. Uma máquina que aprendesse o que fazer em vez de ser programada para o fazer.

A ruptura não estava na complexidade das máquinas. Estava na sua relação com o conhecimento. A máquina de Descartes sabia porque lhe tinham dito. A máquina que as redes neuronais tornaram possível sabe porque aprendeu.

E a diferença entre estas duas frases — entre dizer e aprender, entre programar e treinar, entre o conhecimento como instrução e o conhecimento como emergência — é o abismo conceptual que este livro se propõe atravessar. Nos capítulos que se seguem, vamos explorar como este abismo foi cruzado, o que existe do outro lado e o que a travessia nos revela sobre a natureza da inteligência — a artificial e, inseparavelmente, a nossa.



CAPÍTULO 2

O Fracasso das Regras e a Inteligência Tácita

"Nós sabemos mais do que conseguimos dizer."

Michael Polanyi, *The Tacit Dimension*

I. O Problema que Não Tinha Nome

Em 1966, o Departamento de Defesa dos Estados Unidos encomendou um relatório sobre o estado da tradução automática. Havia quinze anos que o governo americano investia em programas para traduzir automaticamente textos russos — a Guerra Fria tornava o problema urgente e os serviços de informações queriam acesso rápido a documentos científicos e militares soviéticos. Os resultados eram frustrantes. O relatório ALPAC concluiu que a tradução automática era mais lenta, menos precisa e o dobro do preço da tradução humana e que não havia perspectivas de melhoria a curto prazo. O financiamento foi cortado drasticamente.

O que tinha corrido mal? Tecnicamente, tudo parecia razoável. Os linguistas tinham produzido gramáticas formais detalhadas, com centenas de regras que descreviam as estruturas do russo e do inglês. Os programadores tinham implementado essas gramáticas em código. As máquinas podiam analisar frases simples com aceitável precisão. Mas a língua real — com as suas ambiguidades, as suas metáforas, os seus contextos implícitos, as suas referências culturais — resistia obstinadamente à formalização.

A frase inglesa «*Time flies like an arrow; fruit flies like a banana*» era um pesadelo para qualquer analisador sintático. Dependendo de como se interpretasse «*flies*» e «*like*», a frase podia ter pelo menos quatro estruturas gramaticais distintas, cada uma perfeitamente válida do ponto de vista formal: as moscas do tempo gostam de uma seta; o tempo voa como uma seta e as moscas-da-fruta gostam de uma banana; o tempo passa e observa uma seta; e outras combinações igualmente plausíveis para um analisador sem contexto. Um falante nativo desambiguava instantaneamente, sem esforço consciente, guiado por algo que não conseguia articular. A máquina não tinha acesso a esse algo.

O problema não era apenas de ambiguidade sintática. Era de conhecimento do mundo. Para desambiguar «banco» em «sentei-me no banco», o sistema precisava de saber que os bancos financeiros não têm assentos e que os bancos de jardim não têm gestores — um conhecimento tão elementar para um humano adulto que nunca precisou de ser

ensinado explicitamente, mas que estava totalmente ausente dos sistemas de tradução. Para interpretar «Está um tempo horrível hoje» como tentativa de iniciar conversa e não como meteorologia, o sistema precisava de saber que os portugueses falam do tempo quando não têm outro assunto óbvio — uma convenção social que nenhuma gramática codifica.

Este era o problema que não tinha ainda um nome preciso. Era vasto, resistente e parecia aumentar em vez de diminuir à medida que os investigadores trabalhavam nele. Mas estava prestes a receber um nome — e uma formulação filosófica que iluminaria não só as limitações da IA, mas a própria natureza do conhecimento humano.



Michael Polanyi era um químico húngaro que emigrou para Inglaterra nos anos trinta, fugindo ao nazismo e que passou a segunda metade da sua vida a fazer filosofia da ciência. Era um observador agudo das práticas científicas reais — não das versões idealizadas que os manuais descreviam — e o que observava perturbava-o. Os cientistas produziam conhecimento de formas que não correspondiam aos métodos que declaravam seguir. Reconheciam experiências promissoras com um faro que não conseguiam articular. Identificavam anomalias relevantes antes de saberem exactamente porque eram relevantes. Escolhiam problemas férteis por um tipo de julgamento que nenhum critério explícito capturava completamente.

Em 1958, publicou «Personal Knowledge» e em 1966 «The Tacit Dimension» — dois livros que introduziram no vocabulário filosófico um conceito que iria revelar-se central para compreender tanto a inteligência humana como os seus limites artificiais: o conhecimento tácito.

A tese de Polanyi era simples de enunciar e profunda nas suas implicações: *sabemos mais do que conseguimos dizer*. Existe uma dimensão do conhecimento humano — talvez a dimensão mais importante — que não pode ser totalmente articulada em palavras, em regras, em proposições explícitas. É o conhecimento que habita nos dedos do pianista, nos olhos do radiologista, na postura corporal do esquiador. É o conhecimento que guia o julgamento do juiz experiente, a intuição do engenheiro, o instinto do treinador desportivo.

Polanyi chamou-lhe conhecimento tácito para o distinguir do conhecimento explícito — o conhecimento que pode ser dito, escrito, ensinado num manual, codificado numa fórmula. O conhecimento explícito é transmissível por linguagem: posso dizer-te a fórmula da água, a data da Batalha de Alcácer Quibir, as regras do xadrez. O conhecimento tácito é, em grande medida, inefável: pode ser demonstrado, pode ser adquirido por prática prolongada, pode ser reconhecido quando se vê em acção —, mas resiste à tradução verbal

completa.

Para ilustrar a distinção, Polanyi usava um exemplo aparentemente trivial: o reconhecimento de rostos. Reconhecemos um rosto numa multidão, em condições difíceis de iluminação, após anos sem ver a pessoa. Mas não conseguimos descrever como o fazemos. Se alguém nos perguntar como reconhecemos o rosto da nossa mãe, o máximo que conseguimos dizer são generalidades — «os olhos, a forma do rosto, o sorriso» — que não seriam suficientes para que outra pessoa reconhecesse a mesma pessoa. O conhecimento que usamos para o reconhecimento é genuíno e eficaz, mas não está em forma verbal.

A implicação para a inteligência artificial era, como Polanyi percebeu antes que a IA existisse como campo estabelecido, devastadora: se a maior parte do conhecimento humano é tácito e se os sistemas de IA só conseguem trabalhar com conhecimento explícito, então a IA baseada em regras estava condenada a ficar para sempre aquém da inteligência que pretendia simular.

II. Três Retratos do Conhecimento Tácito

Para tornar o abstracto concreto, consideremos três domínios onde o conhecimento tácito se manifesta de forma particularmente clara e onde as tentativas de o formalizar em regras falharam de formas instrutivas.

O primeiro é o reconhecimento de rostos. Um adulto humano reconhece instantaneamente milhares de rostos diferentes — incluindo rostos que não vê há décadas, em condições de iluminação variável, com expressões que nunca viu antes, em ângulos distintos do habitual, mesmo parcialmente tapados. Esta capacidade é tão automática que mal a notamos — é o tipo de competência que só se torna visível quando falha, como acontece com a prosopagnosia, a perturbação neurológica que torna os rostos irreconhecíveis.

Mas tente alguém escrever as regras que permitem distinguir o rosto do João do rosto da Maria. Que propriedades geométricas, que proporções, que texturas, que microexpressões distinguem uma face da outra de forma suficientemente precisa para guiar o reconhecimento? Os investigadores de visão computacional tentaram durante décadas codificar estas regras — com abordagens baseadas em geometria facial, em análise de componentes principais, em comparação de características locais. Os resultados eram competentes em condições controladas e quebravam em condições reais. O conhecimento que um humano usa para reconhecer rostos não existe sob a forma de regras — existe sob a forma de uma competência adquirida por exposição repetida desde os primeiros dias de

vida, que não pode ser analisada nem verbalizada de forma completa.

O segundo é a compreensão da linguagem — não a linguagem dos manuais, mas a linguagem viva com toda a sua riqueza pragmática. Quando lemos «O banco estava cheio de gente», sabemos imediatamente — pelo contexto — se «banco» é uma instituição financeira ou um assento. Quando ouvimos alguém dizer «Está um tempo horrível hoje», percebemos que não é necessariamente uma queixa, mas possivelmente uma tentativa de iniciar conversa. Quando alguém nos diz «Claro, está ótimo», com um tom específico, sabemos que está a dizer exactamente o contrário — e sabemos que sabe que sabemos.

Quando interpretamos a ironia, a metáfora, o subentendido, o sarcasmo, a polidez estratégica, estamos a usar um conhecimento sobre o mundo, sobre as intenções dos falantes, sobre as convenções sociais e culturais, que não está em nenhum dicionário nem em nenhuma gramática. Está em nós, acumulado ao longo de anos de imersão numa cultura e numa língua, inacessível à formulação explícita. Um sistema que conhecesse perfeitamente todas as regras gramaticais e todos os significados dos dicionários ainda falharia sistematicamente nestas dimensões da compreensão — porque as regras e os dicionários não as capturam.

O terceiro é o diagnóstico médico — em particular, o julgamento clínico que os médicos experientes desenvolvem ao longo de anos de prática. Um médico experiente entra num quarto e sabe, antes de perguntar qualquer coisa, que o doente está mal. Não consegue dizer exactamente o quê — talvez a cor da pele, talvez a cadência da respiração, talvez qualquer coisa na postura ou no olhar que o seu sistema nervoso registou antes da consciência o processar. É o que os médicos chamam de «aspecto geral» do doente e é um dos indicadores clínicos mais fiáveis na prática médica, precisamente porque integra uma quantidade enorme de informação subtil de forma holística e imediata.

Nenhum manual descreve como desenvolver esta capacidade. «O doente está mal» não é um diagnóstico que se possa codificar em regras do tipo «se X e Y então Z». É um juízo holístico que emerge de milhares de horas de contacto com doentes em estados diversos — uma espécie de calibração do sistema perceptivo que vai muito além do que qualquer instrução explícita pode transmitir. Adquire-se vendo doentes — muitos doentes, durante muitos anos, com feedback constante sobre se o julgamento estava certo ou errado.



O que é que estes três exemplos têm em comum? Em todos eles, a performance humana é extraordinária e, nas suas formas mais avançadas, essencialmente inigualável pelos sistemas baseados em regras. Em todos eles, a tentativa de articular o conhecimento

subjacente em regras explícitas falha ou produz resultados grotescamente incompletos e em todos eles, o conhecimento foi adquirido por exposição e prática prolongadas, não por instrução formal.

Há um quarto aspecto comum que Polanyi identificou com particular lucidez: em todos eles, tentar verbalizar o conhecimento enquanto se executa a competência degrada o desempenho. O pianista que começa a pensar conscientemente na posição dos dedos durante a execução tropeça. O médico que tenta articular explicitamente cada elemento do «aspecto geral» perde a percepção holística. O reconhecimento de rostos que funciona instantaneamente pode paralisar se tentarmos analisá-lo conscientemente peça por peça. Este fenómeno — a degradação da competência pela atenção focada sobre os seus componentes — é o sinal mais claro de que o conhecimento tácito opera num nível diferente do conhecimento explícito e que os dois interferem em vez de se complementarem.

A inteligência artificial baseada em regras estava, estruturalmente, no modo errado: operava exactamente no nível da atenção consciente e explícita que o conhecimento tácito transcende. Era como tentar ensinar alguém a andar de bicicleta descrevendo verbalmente cada micro-movimento de equilíbrio — não apenas ineficaz, mas contraproducente.

III. Dreyfus e a Crítica Filosófica

Em 1972, um filósofo da Universidade de Berkeley chamado Hubert Dreyfus publicou um livro com o título provocatório «What Computers Can't Do». Era um ataque sistemático aos fundamentos filosóficos da inteligência artificial — e foi recebido com uma hostilidade que dizia muito sobre a insegurança do campo em relação às suas próprias premissas.

Dreyfus não era um engenheiro frustrado nem um tecnofóbico. Era um estudioso sério de Heidegger e Merleau-Ponty e usou as ferramentas da fenomenologia continental para fazer algo que os próprios investigadores de IA raramente faziam: examinar as pressuposições filosóficas por baixo do programa de pesquisa. O que descobriu foi que essas pressuposições tinham três séculos de idade e não tinham sido submetidas a qualquer escrutínio crítico.

O argumento central de Dreyfus era que os investigadores de IA tinham herdado, sem o examinar, o pressuposto cartesiano de que a inteligência é essencialmente simbólica e desincorporada — que a mente funciona representando o mundo em símbolos e aplicando regras a essas representações. Mas a fenomenologia mostrava algo radicalmente diferente: que o ser humano não está no mundo como um processador de símbolos distante. Está no mundo como um corpo que age, que tem práticas, que se move num espaço de significados

que não precisa de ser representado explicitamente porque é vivido directamente.

Heidegger tinha mostrado que o modo primário de ser humano no mundo não é o da contemplação teórica — o sujeito que observa os objectos e os representa mentalmente —, mas o do *envolvimento prático*: o martelo que se usa sem pensar no martelo, a porta que se abre sem analisar o mecanismo da dobradiça, a conversa que se tem sem construir proposições lógicas. A transparência das ferramentas no uso — o facto de que as coisas que dominamos desaparecem da consciência enquanto as usamos — é precisamente o sinal do conhecimento tácito em acção. Só quando o martelo parte, ou quando a porta range, ou quando a conversa fica estranha é que os objectos emergem para a consciência explícita. O modo normal de existência não é a representação; é a imersão.

Merleau-Ponty acrescentou a dimensão corpórea, que Dreyfus considerava decisiva: o conhecimento não está apenas na mente. Está no corpo. O pianista que executa uma peça difícil não está a seguir regras armazenadas algures no córtex pré-frontal e traduzidas em comandos musculares; está a deixar que os dedos se movam segundo uma memória muscular que foi incorporada através de anos de prática e que existe no sistema sensoriomotor de uma forma que não passa pela representação simbólica consciente. O jogador de ténis experiente não calcula a trajectória da bola e decide conscientemente como se mover; move-se e a bola cai onde deve cair, porque o seu corpo aprendeu a fazer o que é necessário sem precisar da mediação da consciência.

A IA baseada em regras pressupunha um modelo cartesiano de cognição: uma mente que representa o mundo em símbolos e aplica regras a essas representações. Dreyfus argumentava que este modelo era errado mesmo como descrição da cognição humana — que a cognição humana começa por baixo, no envolvimento corporal prático com o mundo e que a representação simbólica explícita é uma camada superficial e excepcional, não o modo fundamental de ser no mundo.



A crítica de Dreyfus foi recebida com hostilidade pela comunidade de IA. Marvin Minsky chamou-lhe «o filósofo mais ignorante que alguma vez encontrei». O artigo era descartado como filosofia especulativa sem relevância prática — o trabalho de alguém que não percebia de programação a tentar pronunciar-se sobre um campo que não conhecia.

O tempo deu-lhe razão de uma forma que os seus críticos não anteciparam. Não em todos os pormenores — algumas das suas previsões sobre o que os computadores nunca conseguiriam fazer foram desmentidas —, mas no ponto essencial: as limitações que identificou nos sistemas baseados em regras eram reais, tinham raízes filosóficas profundas

e nenhuma quantidade de regras adicionais as ia resolver. O fracasso da tradução automática, o colapso dos sistemas especialistas fora dos seus domínios restritos, a incapacidade persistente de lidar com linguagem natural em toda a sua riqueza pragmática — tudo isto validou o diagnóstico de Dreyfus, mesmo quando os detalhes da sua análise filosófica podiam ser contestados.

Existe um modelo de aquisição de competências — desenvolvido por Dreyfus em colaboração com o irmão Stuart — que ilumina de forma particular a distância entre o conhecimento explícito e o conhecimento tácito. O modelo descreve cinco etapas de progressão numa competência, do principiante ao especialista.

O principiante segue regras explícitas: «se o semáforo está vermelho, pára; se verde, avança». O comportamento é rígido, lento e requer atenção consciente constante. O avançado começa a reconhecer situações recorrentes e a responder de forma diferenciada — «nesta intersecção específica, com este ângulo de visibilidade, a prudência aconselha abrandar». O competente começa a ver o problema de forma holística, com perspectivas e planos em vez de regras isoladas — planeia a viagem em vez de reagir intersecção a intersecção. O proficiente reage a situações de forma intuitiva, sem cálculo consciente explícito — percebe imediatamente que aquele condutor à frente vai travar bruscamente, antes de haver qualquer sinal explícito disso e o especialista — o condutor experiente, o médico veterano, o mestre de xadrez — simplesmente age. Não analisa, não aplica regras, não consulta princípios. Percebe a situação e faz o que é necessário fazer, de uma forma que não consegue articular completamente.

O que este modelo mostra é que a progressão em qualquer competência é uma progressão do explícito para o tácito, das regras para a intuição, do consciente para o automático. O especialista não é alguém que aprendeu muitas regras. É alguém que transcendeu as regras — que internalizou a competência a um nível onde as regras já não são necessárias nem úteis, onde o conhecimento existe de uma forma que é mais rápida, mais fiável e mais flexível do que qualquer sistema de regras poderia ser.

A inteligência artificial baseada em regras estava, permanentemente, no nível do principiante. Não porque lhe faltassem regras. Mas porque a abordagem tornava impossível a transição para os níveis superiores — a transição que é o ponto central de toda a aprendizagem genuína.

IV. O Impasse e a Promessa

Nos anos oitenta, o optimismo dos pioneiros da IA tinha-se transformado numa frustração crescente. Os sistemas especialistas — os programas mais sofisticados da

abordagem simbólica, como o MYCIN e o XCON — funcionavam bem em domínios muito restritos, mas colapsavam quando tentavam generalizar. A manutenção era pesadíssima: cada nova excepção ao padrão geral exigia novas regras e as novas regras criavam conflitos com as regras existentes. O MYCIN, que funcionava bem para infecções bacterianas, era completamente inútil para qualquer outra tarefa médica — e não havia forma de o estender sem reconstruir praticamente tudo.

O segundo Inverno da IA estava a chegar. O financiamento secou. O entusiasmo arrefeceu. O Japão tinha anunciado em 1982 o ambicioso projecto da «Quinta Geração» de computadores, prometendo máquinas inteligentes baseadas em lógica de predicados até 1992. Dez anos depois, o projecto era amplamente considerado um fracasso. A promessa de máquinas inteligentes parecia ter-se revelado uma ilusão — ou pelo menos um objectivo muito mais distante do que qualquer um tinha imaginado.

Mas nos bastidores, um grupo de investigadores continuava a trabalhar numa abordagem diferente. Não queriam codificar o conhecimento. Queriam construir sistemas que aprendessem. Não queriam programar a inteligência. Queriam criar as condições para que a inteligência emergisse do contacto com os dados.

Tinham a seu favor algo que os seus predecessores não tinham tido: computadores cada vez mais rápidos e, pela primeira vez na história, quantidades significativas de dados digitais — textos, imagens, registos de transacções, dados de sensores — que podiam servir como material de treino. A receita continuava a parecer improvável a muita gente. Mas os ingredientes estavam finalmente a acumular-se.



Há uma ironia profunda nesta história que vale a pena contemplar por um momento.

A inteligência artificial nasceu da convicção de que a inteligência era essencialmente explícita — que pensar era manipular símbolos segundo regras e que portanto uma máquina programada com as regras certas poderia pensar. Esta convicção era herdeira directa do racionalismo cartesiano: a mente como mecanismo de regras, o conhecimento como sistema de proposições articuláveis. Era a versão computacional do sonho de Leibniz: *calculemus*.

O fracasso da IA simbólica não foi um fracasso técnico — não foi simplesmente que os computadores eram demasiado lentos ou que os programadores cometeram erros. Foi uma refutação filosófica. Mostrou que o racionalismo cartesiano — a ideia de que a inteligência é essencialmente um sistema de regras explícitas — está, na melhor das

hipóteses, gravemente incompleto. A maior parte da inteligência humana não é explícita. É tácita. Não segue regras. Reconhece padrões. Opera no modo do envolvimento prático que Heidegger descreveu, não no modo da representação simbólica que Descartes pressupôs.

E a ironia é esta: ao tentar construir inteligência artificial e fracassar, aprendemos algo fundamental sobre a inteligência natural. Aprendemos que Polanyi tinha razão — que sabemos muito mais do que conseguimos dizer e que esse excesso de saber sobre o dizível é precisamente onde reside a competência mais profunda. Aprendemos que Dreyfus tinha razão — que a fenomenologia do corpo e do envolvimento prático captura algo sobre a cognição que o paradigma simbólico ignorava e aprendemos que o empirismo de Hume — a ideia de que o conhecimento emerge da experiência e da associação de padrões, não de princípios inatos ou regras apriorísticas — captura algo essencial sobre como a mente funciona que o racionalismo cartesiano sistematicamente obscurecia.

A IA simbólica falhou. Mas o seu fracasso foi, paradoxalmente, um dos maiores progressos na história da filosofia da mente: demonstrou, de forma empírica e irrefutável, que a mente humana não funciona como um sistema de regras explícitas. Que Descartes, ao traçar a linha entre o mecanismo e a mente, tinha posto do lado da mente não apenas a consciência e a alma, mas também o tipo de conhecimento que nenhum mecanismo de regras consegue capturar.

E deixou em aberto a pergunta que os próximos capítulos vão explorar: se a inteligência não é essencialmente explícita — se a inteligência é essencialmente aprendizagem, reconhecimento de padrões, conhecimento tácito —, então é possível construir uma máquina que aprenda? Uma máquina que desenvolva, através da exposição à experiência, o tipo de conhecimento que resiste à verbalização?

Uma máquina que saiba sem conseguir dizer o que sabe?



CAPÍTULO 3

A Ideia Sobrevivente: Inspirados pelo Cérebro

"Em ciência, podemos dizer coisas que parecem loucas; a longo prazo, podem revelar-se certas."

Geoffrey Hinton, entrevista

I. O Neurónio que Aprende

Em 1943, um neurologista e um matemático publicaram em conjunto um artigo de doze páginas que iria mudar, com um atraso de várias décadas, a história da inteligência artificial.

Warren McCulloch era médico e neurocientista, obcecado com a questão de como o cérebro produz o pensamento. Tinha trabalhado durante anos na tentativa de compreender a arquitectura neurológica da percepção e do raciocínio — não apenas como fenómeno biológico, mas como problema lógico e matemático. Walter Pitts era um prodígio da matemática que tinha fugido de casa aos quinze anos, encontrado abrigo intelectual na Universidade de Chicago e se tornara colaborador informal de alguns dos maiores cientistas da época, incluindo Norbert Wiener, o fundador da cibernética. Juntos, publicaram «A Logical Calculus of the Ideas Immanent in Nervous Activity» — um título árido para uma ideia revolucionária.

O artigo propunha um modelo matemático do neurónio biológico. Um neurónio, simplificado ao essencial, é uma unidade que recebe sinais de outros neurónios através das suas dendrites, soma esses sinais e dispara — envia um sinal pelos seus axónios — se a soma ultrapassar um determinado limiar. É, na sua forma mais abstracta, uma operação de soma e limiar: recebe entradas ponderadas, decide se activa ou não. O modelo de McCulloch e Pitts capturava este comportamento numa equação simples.

McCulloch e Pitts mostraram que redes de neurónios com este comportamento podiam, em princípio, computar qualquer função lógica. Não era apenas uma observação biológica. Era uma afirmação sobre o poder computacional dos sistemas nervosos — e, por extensão, uma sugestão de que a lógica e o pensamento podiam emergir de unidades simples ligadas em rede. Um conjunto suficientemente grande e bem organizado de neurónios artificiais podia, em teoria, realizar qualquer computação que uma Máquina de Turing podia realizar.

A ideia era fascinante. Criava uma ponte entre a neurobiologia e a lógica matemática que ninguém tinha atravessado antes. Mas o modelo de McCulloch e Pitts tinha uma limitação decisiva que os próprios autores reconheciam: as conexões entre neurónios — os pesos que determinavam a importância de cada sinal — eram fixas. A rede não aprendia. Era uma máquina lógica, não um sistema adaptativo. Podia computar qualquer função que lhe fosse programada; não podia descobrir por si própria que funções eram relevantes.

Faltava o mecanismo da aprendizagem.



Donald Hebb, psicólogo canadiano da Universidade McGill, forneceu uma pista fundamental em 1949. No seu livro «The Organization of Behavior», propôs um princípio que ficaria conhecido como a regra de Hebb: quando um neurónio A activa repetidamente o neurónio B, a conexão entre eles fortalece-se. «Células que disparam juntas, ligam-se juntas» — a formulação popular do princípio, que não é de Hebb, mas que captura a essência com razoável fidelidade.

A regra de Hebb era uma teoria da aprendizagem por associação. O cérebro aprende reforçando as conexões que se activam em conjunto — exactamente o tipo de aprendizagem empírica que Hume tinha descrito filosoficamente dois séculos antes, quando falava de associação de ideias como o mecanismo fundamental da cognição. O que Hebb fazia era sugerir um mecanismo neurobiológico concreto para esse processo filosófico: a plasticidade sináptica como implementação física do princípio de associação.

A implicação para as redes artificiais era clara: se os pesos das conexões entre neurónios artificiais pudessem ser ajustados com base na experiência — se a rede pudesse aprender quais as conexões a reforçar e quais a enfraquecer —, então talvez fosse possível construir um sistema que desenvolvesse, por exposição a exemplos, as competências que os sistemas baseados em regras não conseguiam adquirir. A aprendizagem não precisava de ser programada explicitamente; podia emergir do ajuste incremental das conexões em resposta ao contacto com os dados.

O problema era técnico e profundo: como ajustar os pesos de forma sistemática e eficiente, de tal modo que a rede convergisse para um comportamento útil em vez de oscilar ou divergir? A regra de Hebb era biologicamente plausível, mas matematicamente insuficiente para garantir a convergência. Faltava um algoritmo de treino robusto — uma forma de dizer à rede, dado um erro que cometeu, exactamente o que mudar e em quanto.

II. O Perceptrão e o Primeiro Inverno

Frank Rosenblatt era um psicólogo na Cornell University com um talento para o entusiasmo que nem sempre servia bem a sua reputação científica — tinha a rara combinação de genialidade genuína e falta de contenção retórica que torna os cientistas ao mesmo tempo influentes e vulneráveis. Em 1957, desenvolveu o Perceptrão e em 1958 apresentou-o publicamente.

O Perceptrão era um neurónio artificial com um mecanismo de aprendizagem supervisionada: os pesos das suas conexões eram ajustados automaticamente com base nos erros que cometia. Se o Perceptrão classificava incorrectamente uma entrada — se dizia «não» quando a resposta correcta era «sim» —, os pesos eram modificados numa direcção que tornava o erro menos provável na próxima vez que encontrasse uma entrada semelhante. Era simples, elegante e funcionava.

O resultado era genuinamente surpreendente. Um Perceptrão podia aprender a distinguir categorias — separar padrões de dois tipos — sem que ninguém lhe dissesse explicitamente as regras de distinção. Aprendia dos exemplos. A imprensa ficou entusiasmada de uma forma que, em retrospectiva, parece desproporcionada, mas que era compreensível dado o momento. O «New York Times» anunciou que a marinha americana tinha construído um embrião de computador que esperava poder andar, falar, ver, escrever, reproduzir-se e ter consciência da sua existência. Rosenblatt não ajudou a moderar as expectativas.

Durante um breve período, o Perceptrão pareceu ser a chave que a inteligência artificial tinha estado à procura. Aqui estava um sistema que aprendia. Aqui estava, talvez, o mecanismo pelo qual a inteligência poderia emergir de componentes simples.



Em 1969, Marvin Minsky e Seymour Papert publicaram «Perceptrons» — uma análise matemática rigorosa e, nos seus efeitos sobre o campo, devastadora. O resultado central era preciso: um Perceptrão simples não conseguia aprender a resolver problemas que não fossem linearmente separáveis. Na linguagem geométrica da matemática, isto significa que o Perceptrão só conseguia separar categorias que podiam ser divididas por uma linha recta (em duas dimensões) ou por um hiperplano (em dimensões superiores).

O exemplo clássico era o XOR lógico — a função que devolve verdadeiro quando exactamente uma das duas entradas é verdadeira e falso nos outros casos. É uma função tão simples que qualquer criança de escola primária a compreende intuitivamente; não é

linearmente separável e portanto um Perceptrão era matematicamente incapaz de aprender. Se um sistema tão simples o derrotava, como poderia o Perceptrão alguma vez abordar os problemas complexos e não lineares do mundo real?

A implicação parecia ser que as redes neuronais artificiais tinham limitações fundamentais que as tornavam inadequadas para os problemas reais da inteligência artificial. O financiamento para investigação em redes neuronais secou quase completamente. As principais conferências deixaram de aceitar artigos na área. O primeiro Inverno das redes neuronais tinha chegado — e viria a durar mais de uma década.

O que Minsky e Papert tinham provado era, estritamente, correcto para Perceptrões de uma única camada. O que não tinham provado — e este foi o equívoco conceptual que moldou o campo durante anos — era que redes com múltiplas camadas de neurónios tinham as mesmas limitações. Uma rede com camadas intermédias entre a entrada e a saída — o que viria a chamar-se uma rede com camadas ocultas — podia, em princípio, aprender representações internas dos dados que tornavam problemas não lineares tratáveis. A questão era como treinar eficientemente uma rede com múltiplas camadas — como ajustar os pesos de camadas que não tinham acesso directo aos erros da saída. O problema existia. A solução também —, mas ainda não tinha sido formulada de forma geral e eficiente o suficiente para ser amplamente usada.

III. A Sobrevivência na Margem

Durante os anos setenta e início dos oitenta, a investigação em redes neuronais continuou, mas na periferia do campo. Sem financiamento generoso, sem prestígio institucional, sem o interesse das grandes universidades e laboratórios que estavam concentrados na IA simbólica e nos sistemas especialistas. Era uma pequena comunidade de convictos — investigadores que acreditavam que a abordagem estava correcta mesmo quando os resultados eram modestos e o cepticismo era a atitude dominante.

Esta situação tem algo de romanesco que não devemos exagerar — a ciência não funciona por heroísmo individual e muitos investigadores neste período estavam simplesmente a fazer o seu trabalho sem consciência dramática do momento histórico. A maior parte dos artigos que escreveram foi lida por dezenas de pessoas, não por milhares. Os progressos eram incrementais e difíceis de demonstrar de forma convincente. Os subsídios eram escassos e os resultados prometidos raramente correspondiam às expectativas dos financiadores.

Mas a sobrevivência da ideia tem também algo de genuinamente instrutivo sobre como o conhecimento científico avança. A ideia central que mantinha viva a investigação em redes

neurónais era simples e poderosa: o cérebro humano funciona. Funciona extraordinariamente bem — melhor do que qualquer sistema artificial construído até hoje e funciona através de redes de neurónios com conexões ajustáveis pela experiência. Se a natureza, através da evolução, encontrou esta solução para o problema da inteligência adaptativa, talvez valha a pena tentar perceber os seus princípios essenciais e reproduzi-los em substratos artificiais.

Não se tratava de imitação biológica ingénua — os investigadores sabiam que os seus modelos eram grosseiras simplificações do neurónio real, que ignoravam a complexidade da arborização dendrítica, dos diferentes tipos de neurotransmissores, da dinâmica temporal dos potenciais de acção. Tratava-se de uma aposta sobre os princípios computacionais: que a aprendizagem distribuída, as representações emergentes e o ajuste incremental por erro eram princípios mais poderosos do que qualquer sistema de regras explícitas, independentemente do substrato em que fossem implementados.



Em paralelo, alguns investigadores estavam a desenvolver peças que viriam a ser essenciais para o ressurgimento das redes neurónais. Yann LeCun, trabalhando inicialmente em Paris e depois nos Bell Labs nos Estados Unidos, desenvolveu as redes convolucionais — uma arquitectura especialmente adequada para processar dados com estrutura espacial, como imagens. A intuição era biológica: o córtex visual humano processa imagens em camadas hierárquicas, onde cada camada detecta padrões progressivamente mais complexos — arestas simples, seguidas de formas, seguidas de objectos. As redes convolucionais implementavam esta hierarquia artificialmente.

LeCun demonstrou em 1989 que uma rede convolucional podia reconhecer dígitos escritos à mão com uma precisão que superava os métodos anteriores. O sistema foi adoptado pelos bancos americanos para processar cheques — durante anos, uma das redes neurónais com maior escala de implantação real no mundo. Mas o impacto mais amplo no campo foi limitado: o problema dos cheques parecia um nicho, não uma prova do princípio geral.

Em 1986, um artigo na revista «Nature» tinha já mudado parte do panorama. David Rumelhart, Geoffrey Hinton e Ronald Williams publicaram uma formulação geral e eficiente do algoritmo de retropropagação do erro — *backpropagation* —, o método que permitia treinar redes neurónais com múltiplas camadas de forma sistemática.

A ideia central do *backpropagation* é elegante: para ajustar os pesos de camadas intermédias que não têm acesso directo ao erro de saída, propaga-se o erro *para trás* pela

rede, camada a camada, usando o cálculo diferencial para determinar qual a contribuição de cada peso para o erro final. É um método de atribuição de responsabilidade distribuída: dada uma resposta errada, quanto é que cada conexão da rede contribuiu para esse erro? Cada conexão é ajustada proporcionalmente à sua responsabilidade — um pouco mais ou um pouco menos, numa direcção que tornará o erro menos provável na próxima vez.

Com o backpropagation, redes com camadas ocultas podiam ser treinadas eficientemente e podiam aprender representações internas dos dados — podiam aprender a ver o mundo de formas que não estavam programadas, que emergiam do próprio processo de aprendizagem. A ferramenta fundamental estava disponível.

Mas o segundo Inverno estava ainda por vir.

IV. O Segundo Inverno e a Teimosia de Hinton

Os anos noventa trouxeram uma nova onda de cepticismo que foi, em alguns aspectos, mais difícil de suportar do que o primeiro Inverno. Desta vez, o problema não era a ausência de resultados — havia resultados. Era que outros métodos pareciam funcionar melhor.

As máquinas de vectores de suporte — um método alternativo desenvolvido por Vladimir Vapnik e Corinna Cortes, com sólidas garantias teóricas provenientes da teoria estatística da aprendizagem — superavam as redes neuronais em muitas tarefas de classificação padrão, especialmente quando os dados eram escassos. Os métodos baseados em árvores de decisão e os seus agregados (florestas aleatórias, gradient boosting) eram robustos, interpretáveis e computacionalmente eficientes. O campo da aprendizagem automática estava a fragmentar-se numa diversidade de abordagens e as redes neuronais profundas perdiam terreno para métodos que pareciam mais sólidos teoricamente e mais fáceis de treinar na prática.

Geoffrey Hinton, que tinha sido um dos autores do artigo de 1986 sobre backpropagation, recusou-se a desistir. Era uma posição intelectualmente solitária e profissionalmente arriscada. Enquanto a maioria dos investigadores migrava para outros métodos ou trabalhava com redes rasas de uma ou duas camadas, Hinton continuou a trabalhar em redes neuronais profundas — redes com muitas camadas ocultas. Acreditava que a profundidade era essencial: que as representações hierárquicas aprendidas por redes profundas capturariam algo sobre a estrutura multinível do mundo que redes rasas não conseguiam capturar.

O problema era prático e matemático. Redes profundas eram difíceis de treinar com backpropagation. O gradiente — o sinal que guia o ajuste dos pesos — tendia a desaparecer nas camadas mais afastadas da saída, tornando-as impossíveis de treinar eficientemente; ou,

pelo contrário, tendia a explodir, produzindo actualizações desestabilizadoras. A profundidade que Hinton queria estava, por razões técnicas que não eram imediatamente óbvias como resolver, fora de alcance prático.

Durante anos, o trabalho de Hinton e dos seus colaboradores foi recebido com indiferença ou hostilidade pelas principais conferências da área. Propostas de artigos eram rejeitadas com comentários que sugeriam que a abordagem era um beco sem saída. Financiamento era difícil de obter para trabalho em redes neuronais profundas quando os revisores acreditavam que o problema estava resolvido por outros métodos. Havia uma sensação generalizada — não totalmente irracional dado o estado das evidências — de que as redes neuronais tinham tido a sua oportunidade e não tinham correspondido às expectativas.

Hinton continuou. Yann LeCun, que continuava a trabalhar em arquitecturas convolucionais nos Bell Labs e depois na New York University, continuou. Yoshua Bengio, em Montreal, que se tinha especializado em aprendizagem de representações de linguagem, continuou. Estes três investigadores — que viriam a ser conhecidos informalmente como a «máfia do deep learning» e que em 2018 receberiam conjuntamente o Prémio Turing, o equivalente do Nobel em ciência da computação — partilhavam uma convicção que os dados ainda não suportavam plenamente, mas que os mantinha a trabalhar: as redes profundas eram o caminho certo.



O que os mantinha? Esta é uma pergunta que vale a pena fazer, não por romantismo histórico ou para construir uma narrativa heróica, mas porque a resposta ilumina algo sobre como o conhecimento científico avança — e sobre a diferença entre seguir a evidência disponível e apostar numa ideia cujos frutos ainda não amadureceram.

Parte da resposta é técnica: Hinton, LeCun e Bengio tinham razões matemáticas e empíricas para acreditar que a abordagem era correcta, mesmo quando os resultados práticos eram decepcionantes. Os problemas que encontravam pareciam-lhes problemas de recursos — de computação insuficiente, de dados insuficientes, de técnicas de treino imaturas — e não problemas de princípio. Havia resultados parciais suficientemente encorajadores para sustentar a convicção de que, com mais recursos, a abordagem funcionaria.

Parte da resposta é filosófica: a convicção de que o cérebro — que funciona — usava estes princípios e que portanto os princípios tinham de ser viáveis em algum substrato. Era uma forma de argumento naturalista que não é logicamente necessário, mas que é

heurísticamente poderoso: a natureza já resolveu o problema da inteligência adaptativa; o que precisamos de perceber é como, não se é possível. As redes neuronais profundas eram a melhor hipótese disponível sobre o «como».

E parte da resposta é, provavelmente, simplesmente a combinação de teimosia e auto-confiança que distingue o investigador que faz história do investigador que segue a corrente. Hinton diria mais tarde, com algum humor seco, que a vantagem de não ter uma reputação a perder num domínio específico é que se pode continuar a trabalhar em ideias impopulares sem o custo de reputação que afastaria outros. Mas havia algo mais do que teimosia — havia uma intuição profunda, alimentada por anos de trabalho próximo com o problema, de que as redes profundas eram genuinamente diferentes em capacidade das redes rasas e que essa diferença importaria quando os recursos fossem suficientes.

V. A Chegada do Dilúvio

Em 2012, uma equipa liderada por Hinton e dois dos seus estudantes de doutoramento — Alex Krizhevsky e Ilya Sutskever — submeteu um sistema ao ImageNet Large Scale Visual Recognition Challenge, uma competição anual de reconhecimento de imagem que tinha emergido como o benchmark do estado da arte na visão computacional. O ImageNet era uma base de dados de mais de um milhão de imagens classificadas em mil categorias — cães de raças específicas, espécies de pássaros, tipos de instrumentos musicais, marcas de automóveis — e a tarefa era classificar correctamente imagens nunca vistas.

O sistema chamava-se AlexNet. Usava uma rede neuronal profunda com múltiplas camadas convolucionais, em parte inspirada no trabalho de LeCun, mas treinada numa escala sem precedentes em unidades de processamento gráfico — as GPUs originalmente desenvolvidas para jogos de vídeo, que se revelaram excepcionalmente adequadas para os cálculos matriciais que as redes neuronais requerem. A combinação de arquitectura profunda, dados massivos e GPUs produziu algo que ninguém esperava.

Os resultados foram perturbadores. A taxa de erro do AlexNet no teste de reconhecimento foi de 15,3%, contra 26,2% do segundo classificado — uma melhoria de mais de dez pontos percentuais numa competição onde melhorias de um ou dois pontos eram consideradas significativas. Não era uma melhoria incremental. Era uma ruptura de magnitude que o campo não tinha visto.

A comunidade de visão computacional ficou em choque. Em doze meses, quase todos os grupos de investigação relevantes tinham abandonado as abordagens anteriores — os métodos baseados em características desenhadas manualmente, as máquinas de vectores de suporte aplicadas a representações cuidadosamente construídas — e estavam a trabalhar

com redes neuronais profundas. O investimento que tinha sido recusado durante anos convergiu de repente, de empresas tecnológicas, de governos, de fundos de capital de risco. O Inverno tinha acabado e o fim não foi gradual — foi uma avalanche.



Nos anos seguintes, o padrão repetiu-se em domínio após domínio com uma velocidade e uma consistência que desafiavam a previsão. Reconhecimento de voz: os sistemas baseados em redes neuronais profundas reduziram a taxa de erro pela metade em poucos anos. Tradução automática: os modelos neuronais superaram os melhores sistemas baseados em frases e regras linguísticas com margens que os próprios investigadores consideraram implausíveis até as medir. Jogos de tabuleiro: o AlphaGo da DeepMind derrotou o campeão mundial de Go em 2016, num jogo considerado demasiado complexo e intuitivo para as máquinas. Geração de texto, diagnóstico médico, síntese de imagem — em cada caso, as redes neuronais profundas, treinadas em quantidades massivas de dados com poder computacional crescente, superaram tudo o que tinha existido antes.

O que tinha mudado? Tecnicamente, três coisas tinham convergido de forma que ninguém tinha antecipado completamente. Primeiro, as arquitecturas melhoraram — novas formas de organizar as camadas, novas funções de activação, novas técnicas de regularização que tornavam o treino mais estável e eficiente. Segundo, os dados digitais tornaram-se abundantes: a internet tinha produzido, ao longo de uma década, quantidades imensas de texto, imagem, vídeo e áudio que podiam servir como material de treino. Terceiro, o poder de computação aumentou exponencialmente, sobretudo através das GPUs que permitiam paralelizar o treino de forma que os CPUs tradicionais não podiam.

Mas por baixo das inovações técnicas havia algo mais fundamental: a confirmação de que a aposta filosófica estava correcta. Que a aprendizagem era mais poderosa do que a programação. Que as representações emergentes — as que a rede aprende por si própria a partir dos dados — eram mais ricas e mais eficazes do que as representações desenhadas manualmente por engenheiros humanos. Que o conhecimento tácito que Polanyi descreveu — o conhecimento que emerge da exposição a exemplos e que não pode ser completamente articulado em regras — podia ser, pelo menos parcialmente, capturado por sistemas artificiais.

A ideia que tinha sobrevivido ao Inverno durante décadas, mantida viva por um grupo pequeno de investigadores contra o ceticismo dominante e a escassez de recursos, tinha provado o seu valor de uma forma que o campo nunca mais poderia ignorar.

Mas o que, exactamente, tinha provado? Tinha provado que as redes profundas são

úteis — isso era evidente. Tinha provado que a aprendizagem a partir de dados é mais poderosa do que a programação de regras — também evidente. O que não era evidente e o que os capítulos seguintes vão explorar, é o que esta prova nos diz sobre a natureza do conhecimento, da inteligência e da mente. Porque a ruptura de 2012 não foi apenas uma ruptura técnica. Foi uma ruptura filosófica — a demonstração empírica de que certas ideias sobre a natureza da inteligência que tinham sido consideradas marginais ou especulativas eram, afinal, correctas.

É para estas perguntas — o que provou o sucesso das redes neuronais e o que nos diz sobre nós próprios — que nos voltamos agora.



PARTE II

A Descoberta Fundamental: Ensinar a Aprender

CAPÍTULO 4

Inversão de Paradigma: O Avanço Não Foi Pensar, Foi Aprender

"Em vez de tentar produzir um programa que simule a mente adulta, porque não tentar produzir um que simule a mente de uma criança?"

Alan Turing, Computing Machinery and Intelligence, 1950

I. A Distinção que Tudo Muda

Existe uma distinção que parece óbvia depois de ser dita — e que, precisamente por isso, levou décadas a ser reconhecida na sua plena profundidade. É uma dessas distinções que, quando finalmente se torna clara, transforma o passado: ilumina retroactivamente porque certas abordagens falharam, porque certas intuições resistiram quando deviam ter sido abandonadas, porque o campo progrediu de formas que nenhum dos seus fundadores antecipou.

Pensar e aprender não são a mesma coisa.

Pensar — no sentido em que os pioneiros da inteligência artificial usavam a palavra — é aplicar regras a representações. É a operação que Descartes tinha em mente quando falava da mente como *res cogitans*, a coisa que pensa. É o que Leibniz sonhava automatizar com o seu *calculus ratiocinator*. É o que os sistemas especialistas dos anos oitenta tentavam fazer: dado um conjunto de factos e um conjunto de regras, derivar conclusões. O pensamento, neste modelo, é um processo dedutivo: parte-se do geral para o particular, das premissas para as conclusões, das regras para os casos.

Aprender é diferente em natureza, não apenas em grau. Aprender é modificar o próprio sistema com base na experiência, de forma a tornar as respostas futuras melhores do que as passadas. Não é aplicar regras. É ajustar os parâmetros internos do sistema — os pesos, as representações, as tendências — em função do contacto com o mundo. É um processo indutivo: parte-se dos casos particulares para as generalizações, da experiência acumulada para o padrão emergente.

A distinção parece simples. Mas as suas implicações são vertiginosas e a primeira implicação é sobre o estatuto do conhecimento em cada um dos dois tipos de sistema.

Um sistema que pensa precisa que alguém lhe diga o que saber. O conhecimento tem

de ser introduzido explicitamente — codificado em regras, inscrito em bases de dados, programado por engenheiros que entendem o domínio. O sistema é, em última análise, um executor sofisticado do conhecimento dos seus criadores. É *derivativo*: o seu conhecimento é sempre uma cópia, mais ou menos elaborada, do conhecimento humano que lhe foi fornecido. Se o engenheiro soube, o sistema sabe. Se o engenheiro não soube — ou não conseguiu articular o que sabia —, o sistema não sabe.

Um sistema que aprende adquire conhecimento por si próprio — não por instrução explícita, mas por exposição a exemplos, por ajuste incremental em resposta ao erro, por abstracção de padrões que emergem dos dados. O seu conhecimento não é uma cópia do que lhe foi dito. É algo que *emerge* do processo de aprendizagem — algo que, em muitos casos, nem os seus criadores conseguem articular completamente. O sistema pode saber coisas que nenhum dos seus criadores sabia dizer. Pode desenvolver competências que ninguém programou explicitamente para ele ter.

Esta diferença — entre o sistema que recebe conhecimento e o sistema que desenvolve conhecimento — é a ruptura conceptual central da IA moderna. É o que torna as redes neuronais profundas fundamentalmente diferentes não apenas dos sistemas especialistas, mas de toda a história prévia da tecnologia: pela primeira vez, construímos artefactos cujo comportamento e cujo conhecimento não foram completamente especificados pelos seus criadores.



Há uma forma de tornar esta distinção mais concreta que recorre a um exemplo familiar: a aprendizagem de uma língua.

Um tradutor automático baseado em regras — do tipo que dominou o campo até meados dos anos dois mil — funciona assim: linguistas analisam as gramáticas de duas línguas, identificam as correspondências entre estruturas, codificam as regras de transformação num programa. O sistema recebe uma frase numa língua, analisa-a segundo as regras da sua gramática, constrói uma representação intermédia abstracta e aplica as regras de transformação para gerar a frase equivalente na outra língua.

O processo é transparente e auditável. Cada passo pode ser examinado. Cada erro pode ser rastreado à regra que o produziu. O engenheiro pode corrigir a regra e o erro desaparece. O sistema é completamente explicável: dado qualquer output, é possível reconstruir o raciocínio que o produziu. O conhecimento que o sistema usa foi introduzido pelos linguistas. Se eles erraram uma regra, o sistema erra da mesma forma. Se a língua tem uma construção que as regras não cobrem, o sistema falha de forma previsível. O sistema não

sabe nada que os linguistas não lhe tenham ensinado — e não pode saber.

Um tradutor automático baseado em redes neuronais funciona inteiramente de outra forma. Mostra-se ao sistema centenas de milhões de pares de frases — uma frase em português e a sua tradução em inglês, uma frase em francês e a sua tradução em japonês, repetido até à exaustão. O sistema ajusta os seus milhares de milhões de parâmetros internos até ser capaz de produzir traduções que correspondem aos exemplos que viu e que generalizam correctamente para frases que nunca viu. Ninguém lhe explicou a gramática do português nem a do inglês. Ninguém lhe descreveu as regras de concordância verbal, a posição dos adjectivos, as construções idiomáticas, as diferenças de registo entre o formal e o coloquial. O sistema extraiu tudo isso dos dados — de forma implícita, tácita, codificada em pesos numéricos que não têm tradução em regras que alguém possa escrever num manual.

E o resultado é dramaticamente superior ao sistema baseado em regras. Não porque o sistema seja mais inteligente no sentido de raciocinar melhor. Mas porque o conhecimento que adquiriu por aprendizagem é mais rico, mais matizado, mais sensível ao contexto, mais capaz de lidar com a ambiguidade e a variedade da língua real do que qualquer conjunto de regras que linguistas humanos conseguissem formular. O sistema sabe mais do que foi ensinado — precisamente porque o que aprendeu não é o que lhe foi ensinado, mas o que emergiu do contacto com os dados.

Este exemplo da tradução automática é instrutivo porque a história é conhecida e os resultados são mensuráveis. Mas o mesmo padrão — a superioridade do conhecimento emergente sobre o conhecimento programado — repetiu-se em domínio após domínio: reconhecimento de voz, diagnóstico médico, previsão de estruturas proteicas, geração de imagens, raciocínio sobre texto. Em cada caso, o sistema que aprende supera o sistema que pensa.

II. De Cima para Baixo, de Baixo para Cima

Os investigadores em inteligência artificial usam frequentemente dois termos para descrever as abordagens opostas: top-down e bottom-up — de cima para baixo e de baixo para cima. A terminologia é técnica, mas o que descreve é filosófico.

A IA simbólica é top-down: começa com os princípios gerais, as regras, as estruturas abstractas que os especialistas humanos identificam e desce até aos dados concretos. O conhecimento flui do geral para o particular. O especialista humano define as categorias e as relações; o sistema aplica-as aos casos. A arquitectura do conhecimento é determinada antes de o sistema ver qualquer dado.

A aprendizagem automática — e em particular as redes neuronais profundas — é *bottom-up*: começa com os dados concretos, os exemplos brutos, os padrões particulares e ascende até às estruturas abstractas que emergem da regularidade dos dados. O conhecimento emerge do particular para o geral. Ninguém define as categorias; elas formam-se pelo agrupamento espontâneo de exemplos similares. A arquitectura do conhecimento emerge depois de o sistema ver os dados — e não podia ter sido prevista antes.

Esta inversão não é apenas técnica. É filosófica no sentido mais profundo — reflecte uma escolha entre duas epistemologias que a filosofia ocidental debateu durante séculos.

A abordagem *top-down* pressupõe que as estruturas correctas já são conhecidas — que o trabalho é codificá-las e aplicá-las. É uma epistemologia racionalista: o conhecimento começa com princípios e desce até à experiência. É, no fundo, a epistemologia de Descartes, que acreditava que as verdades fundamentais podiam ser conhecidas pela razão *a priori*, independentemente da experiência. É também a epistemologia de Kant, que argumentava que certas categorias do entendimento — causalidade, substância, espaço, tempo — são estruturas que a mente impõe à experiência em vez de extrair dela. O racionalismo, nas suas várias formas, parte sempre da premissa de que há algo que o pensamento sabe antes de olhar para os dados.

A abordagem *bottom-up* pressupõe que as estruturas correctas ainda não são conhecidas — que têm de ser descobertas a partir dos dados. É uma epistemologia empirista: o conhecimento começa com a experiência e ascende até às generalizações. É a epistemologia de Locke — a mente como *Tabula Rasa*, sem ideias inatas, que adquire todo o conhecimento através dos sentidos. É a epistemologia de Hume, que argumentava que todas as nossas ideias, por mais abstractas que pareçam, têm a sua origem em impressões sensoriais concretas e que as generalizações que fazemos são o produto da associação habitual de padrões — não de deduções a partir de princípios inatos. O empirismo, nas suas várias formas, parte sempre da premissa de que o conhecimento é posterior à experiência, não anterior a ela.

As redes neuronais são, neste sentido, a materialização computacional do empirismo filosófico. São a *Tabula Rasa* de Locke tornada máquina: um sistema que começa sem conhecimento inato — apenas com uma arquitectura que define a forma como pode aprender — e que adquire todo o seu conhecimento através da experiência com os dados. A diferença crucial é que a «experiência» de uma rede neuronal pode ser comprimida em horas de treino o que um humano levaria décadas a acumular: uma rede treinada em cem milhões de frases viu mais linguagem do que qualquer humano alguma vez leu.



É importante, aqui, introduzir uma ressalva que o leitor exigente já está provavelmente a formular — e introduzi-la com a seriedade que merece, não como codicilo burocrático, mas como distinção filosoficamente relevante.

A analogia entre redes neuronais e empirismo filosófico é uma ferramenta conceptual, não uma afirmação biológica. Não estamos a dizer que o cérebro humano funciona como uma rede neuronal artificial — que os neurónios biológicos são matematicamente equivalentes aos nós de uma rede de aprendizagem profunda, ou que a aprendizagem humana é o mesmo processo que o backpropagation.

A primeira diferença importante é a eficiência da amostra. Uma criança humana aprende o conceito de «gato» vendo dois ou três gatos, em contextos variados, ouvindo o nome dito por adultos em situações específicas. Uma rede neuronal precisa de dezenas ou centenas de milhares de imagens de gatos para aprender a categoria com fiabilidade comparável. Esta assimetria não é trivial: sugere que a mente humana não é puramente empirista no sentido radical de Hume. Tem uma predisposição genética, um conjunto de estruturas inatas que orientam a aprendizagem e a tornam dramaticamente mais eficiente do que qualquer processo puramente indutivo a partir de dados brutos. Chomsky chamou a isto a «gramática universal»; os psicólogos evolutivos falam de módulos cognitivos — predisposições especializadas para aprender certas coisas (linguagem, faces, agentes causais, objectos permanentes) de forma muito mais rápida do que seria possível por indução geral.

A segunda diferença fundamental é a corporeidade. Nós aprendemos com o corpo inteiro: a dor ensina-nos a evitar o perigo antes de podermos articular o conceito de perigo, a fome ensina-nos a valorizar a comida antes de compreendermos o que é metabolismo, o prazer do toque ensina-nos vínculos afectivos antes de termos linguagem para os nomear. Aprendemos interagindo fisicamente com um mundo que resiste, que surpreende, que tem consequências reais para organismos com necessidades reais. Uma rede neuronal aprende num vácuo de matrizes matemáticas, sem peso, sem fome, sem dor, sem história evolutiva, sem o corpo que ancora a experiência humana ao mundo com uma urgência que nenhum gradiente de optimização consegue replicar.

Estas ressalvas são essenciais. Mas não invalidam a analogia — transformam-na no que deve ser: não uma identidade entre a mente humana e a rede neuronal, mas uma prova de conceito de que o princípio empirista é computacionalmente viável. Que é possível construir um sistema que adquira conhecimento por exposição a exemplos sem que esse conhecimento tenha sido explicitamente programado. Esta é uma afirmação filosófica importante que não precisava de uma mente humana para ser demonstrada — precisava de

uma rede neuronal.

III. O que é uma Representação

Uma das contribuições mais importantes das redes neuronais profundas — e uma das menos discutidas fora dos círculos técnicos, talvez porque exige alguma familiaridade com álgebra linear para ser descrita com precisão — é o que fazem com a noção de representação.

Nos sistemas de IA tradicionais, as representações eram desenhadas à mão por engenheiros. Para reconhecer rostos, um investigador de visão computacional identificava as características relevantes — a distância entre os olhos, a proporção do nariz relativamente à boca, a curvatura das sobrancelhas, a textura da pele — e escrevia código que extraía estas características de cada imagem e as combinava num vector numérico que representava o rosto. O sistema classificava depois os rostos com base neste vector de características desenhado manualmente.

O problema era que as características que os engenheiros escolhiam eram as características que eles conseguiam articular consciente e explicitamente — as que pareciam relevantes de um ponto de vista humano e o que o capítulo anterior mostrou é que o conhecimento humano sobre como reconhecer rostos é em grande medida tácito: não corresponde a uma lista articulável de características geométricas, mas a um padrão holístico de reconhecimento que não pode ser completamente traduzido em variáveis mensuráveis.

Nas redes neuronais profundas, as representações não são desenhadas. Emergem do próprio processo de aprendizagem. As camadas iniciais da rede aprendem a detectar padrões muito simples nos dados — no caso das imagens, arestas orientadas em diferentes direcções, gradientes locais de intensidade, pequenas texturas. Estas detectoras de características primitivas não foram programadas; formaram-se porque eram úteis para reduzir o erro de classificação.

As camadas intermédias combinam os padrões das camadas anteriores em estruturas progressivamente mais complexas — combinações de arestas que formam cantos, combinações de cantos que formam formas, combinações de formas que formam partes de objectos. Cada camada aprende a ver o mundo a um nível de abstracção superior ao da camada anterior e as camadas mais profundas aprendem representações de alto nível que correspondem ao que poderíamos chamar conceitos — «este padrão é um rosto», «esta configuração sugere alegria», «este objecto é um gato» — sem que ninguém tenha definido o que estes conceitos significam em termos das características de baixo nível.

Este processo de aprendizagem hierárquica de representações é, talvez, o aspecto mais

surpreendente e filosoficamente rico das redes neuronais modernas. As representações que emergem não são as que os engenheiros teriam escolhido — são frequentemente mais eficazes, porque capturam aspectos dos dados que os humanos não tinham identificado conscientemente como relevantes e são, em grande medida, ininteligíveis para os humanos: se tentarmos visualizar o que uma camada intermédia de uma rede profunda «vê» quando processa uma imagem, obtemos padrões que activam maximamente aquela camada, mas que não correspondem a nenhuma categoria que o nosso sistema visual reconheça como significativa. São como a linguagem de um sistema perceptivo que evoluiu de forma diferente do nosso.

O conhecimento da rede existe — é real, operacional, eficaz. As representações que usa funcionam melhor do que as representações desenhadas manualmente pelos engenheiros. Mas estão codificadas numa linguagem — a linguagem dos pesos numéricos, das activações, das geometrias em espaços de centenas ou milhares de dimensões — que não tem tradução directa para o português ou para qualquer língua que os humanos falem.

Voltámos a Polanyi, mas desta vez a partir de um ângulo diferente. O médico que reconhece o diagnóstico e o mestre de xadrez que vê a jogada certa têm conhecimento tácito que não conseguem verbalizar. Isso era filosoficamente interessante, mas podia ser tratado como uma limitação contingente: talvez, com mais introspecção ou com melhores técnicas de explicitação, o conhecimento pudesse ser articulado. As redes neuronais fecham esta saída: o seu conhecimento não é apenas difícil de articular — existe numa forma que é estruturalmente irreduzível à linguagem. Não é uma questão de esforço insuficiente. É uma questão de que o tipo de conhecimento que as redes desenvolvem não tem, pela sua própria natureza, uma representação proposicional.

IV. A Escala como Revelação

Nos últimos anos, a investigação em IA descobriu algo que era imprevisível e que continua a ser debatido com intensidade: que o aumento da escala — mais dados, mais parâmetros, mais computação — produz não apenas melhorias quantitativas no desempenho, mas por vezes saltos qualitativos em capacidade.

Os investigadores chamam a isto «capacidades emergentes»: competências que os modelos grandes exibem e que os modelos pequenos não exibem, mesmo que os modelos grandes sejam apenas versões numericamente ampliadas da mesma arquitectura com os mesmos princípios de treino. Um modelo com cem mil milhões de parâmetros faz coisas que um modelo com mil milhões de parâmetros não faz — não apenas melhor, mas categoricamente diferentes, numa descontinuidade que não era previsível pela extrapolação

linear da curva de desempenho. Raciocínio analógico complexo. Resolução de problemas de lógica multi-passo. Geração de código funcional. Seguimento de instruções abstractas nunca vistas. Transferência de conhecimento entre domínios aparentemente não relacionados.

Isto é desconcertante por uma razão precisa: se o sistema é apenas maior — mais do mesmo algoritmo aplicado a mais dados —, porque produz capacidades que não estavam presentes na versão menor? A resposta mais honesta é que não sabemos completamente. A hipótese mais plausível, desenvolvida por investigadores como Wei et al. na Google, é que certas capacidades requerem representações internas de uma riqueza e de uma organização que só são atingíveis acima de um certo limiar de escala — que abaixo desse limiar a rede não tem «espaço representacional» suficiente para desenvolver as estruturas de que precisaria para executar a tarefa. A capacidade é, neste sentido, similar a uma mudança de fase: a água permanece líquida até 100 graus e depois passa abruptamente a vapor. Não há água «um pouco gasosa» a 99 graus.

Seja qual for a explicação técnica precisa, o fenómeno tem implicações filosóficas que vão além da engenharia. Sugere que a inteligência — ou pelo menos o comportamento inteligente — pode não ser apenas uma questão de algoritmo. Pode ser também uma questão de escala. O que emerge a grande escala pode ser qualitativamente diferente do que existe a pequena escala, mesmo com os mesmos princípios fundamentais. A quantidade transforma-se em qualidade — o que é uma ideia que o pensamento filosófico reconhece, mas que raramente tinha sido demonstrada de forma tão concreta.

Isto não é estranho à biologia. Um neurónio não pensa. Dois neurónios não pensam. Cem mil milhões de neurónios, ligados de formas específicas e treinados pela experiência de uma vida inteira, produzem a experiência consciente — ou algo que se lhe assemelha tanto que a distinção entre «produzir» e «ser» se torna filosófica. A emergência de propriedades genuinamente novas a partir de componentes simples em quantidade e organização suficientes é um dos padrões mais recorrentes na natureza, dos cristais às colónias de formigas ao sistema nervoso.

As redes neuronais artificiais estão a reproduzir, em escala muito menor e com mecanismos muito diferentes, algo que ressoa com este padrão de emergência e isso levanta questões que vão muito além da engenharia — questões sobre a natureza da inteligência, sobre a relação entre complexidade e capacidade, sobre o que significa que um sistema desenvolva propriedades que não estavam programadas nos seus componentes.



Chegámos ao ponto em que a distinção entre pensar e aprender — com que abrimos este capítulo — se mostra na sua plena profundidade e nas suas implicações mais perturbadoras.

Os sistemas que aprendem não são apenas sistemas que pensam de forma diferente. São sistemas que adquirem conhecimento de uma forma fundamentalmente diferente — uma forma que os torna, em certos aspectos, mais semelhantes a entidades que crescem e se desenvolvem do que a máquinas que executam programas. O seu conhecimento não foi especificado pelo criador; emergiu do contacto com os dados. As suas capacidades não foram programadas; formaram-se durante o treino. Os seus limites não são os limites das regras que alguém escreveu; são os limites do que pode ser aprendido a partir dos dados disponíveis.

Não são programados para reconhecer rostos — desenvolvem essa capacidade pela exposição a imagens. Não são programados para traduzir — aprendem a fazê-lo pela exposição a pares de frases. Não são programados para raciocinar por analogia — esta capacidade emerge, a certa escala, do processo de aprendizagem sobre texto que contém raciocínio analógico humano.

E aqui está o que é verdadeiramente perturbador e que abre o território do próximo capítulo: nenhum dos seus criadores sabe completamente *como* estas capacidades emergem. Sabem as condições em que emergem — os dados, a arquitectura, a escala, o algoritmo de treino. Mas o processo interno pelo qual um conjunto de pesos numéricos se transforma em competência de tradução, em diagnóstico médico, em raciocínio analógico, é, em grande medida, opaco. Não apenas difícil de descrever — genuinamente opaco, no sentido de que não existe uma descrição em linguagem humana que corresponda ao que acontece dentro da rede quando ela processa uma entrada.

O sistema sabe. Os criadores não sabem completamente o que o sistema sabe nem como sabe. Esta é uma situação sem precedente na história da tecnologia — e é o paradoxo que o próximo capítulo vai explorar.



CAPÍTULO 5

O Regresso da Caixa Negra e a Ruptura Epistemológica

"Os grandes cientistas são homens de ideias ousadas, mas altamente críticos das suas próprias ideias."

Karl Popper, *The Problem of Demarcation*

I. A Exigência de Transparência

Em 1628, William Harvey publicou uma das obras mais importantes da história da medicina: «*Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus*» — um tratado sobre o movimento do coração e do sangue nos animais. Harvey demonstrava, através de observação cuidadosa e argumentação rigorosa, que o sangue circula no corpo numa direcção constante, impulsionado pelo coração como por uma bomba. Era uma tese que contrariava séculos de medicina galénica, que acreditava que o sangue era produzido continuamente no fígado e consumido pelos tecidos.

O que tornava a descoberta de Harvey revolucionária não era apenas o seu conteúdo — era o seu método. Harvey não se limitava a descrever o que observava. Explicava o mecanismo. Mostrava como funcionava através de medições e experiências que outros podiam repetir. Calculou que a quantidade de sangue que o coração bombeia por hora é muito superior à que o fígado poderia produzir, tornando impossível a teoria galénica. Fornecia um modelo causal verificável, não apenas uma narrativa plausível. A transparência do mecanismo era parte integrante do conhecimento — e era o que distinguia a ciência da autoridade.

Este princípio — que conhecer algo significa compreender o mecanismo pelo qual funciona — tornou-se um dos pilares da ciência moderna. Não basta observar que A se segue a B. É preciso perceber porquê, de forma suficientemente precisa para que a previsão seja possível, para que o erro seja identificável, para que outros possam verificar o raciocínio. A caixa negra — o sistema que produz resultados sem revelar o processo — pode ser útil na prática, mas não é ciência no sentido pleno. A ciência exige abertura do mecanismo à análise e à crítica.

As redes neuronais profundas são, neste sentido, uma provocação directa ao programa epistemológico que Harvey ajudou a fundar. Funcionam extraordinariamente bem — melhor, em muitos domínios, do que qualquer sistema que a ciência construiu antes e são,

em grande medida, caixas negras. O paradoxo é esta coexistência: desempenho superior sem explicação correspondente.



O problema não é de princípio — não é que as redes neurais escondam intencionalmente o seu funcionamento, ou que os seus criadores sejam relutantes em explicar. É um problema de escala e de natureza que não tem precedente na história da tecnologia.

Uma rede neuronal com mil milhões de parâmetros é, tecnicamente, completamente determinística e completamente descrita pelas suas equações matemáticas. Nada está escondido no sentido de que algo está inacessível em princípio: os pesos estão todos registados em ficheiros que qualquer um pode ler, as operações são todas conhecidas — multiplicações de matrizes, aplicações de funções não lineares —, o processo pelo qual uma entrada se transforma numa saída pode ser traçado passo a passo com exactidão matemática perfeita.

Mas esta transparência formal é quase completamente inútil do ponto de vista do entendimento. Dizer que uma rede neuronal que diagnostica cancro é uma composição de dois mil milhões de multiplicações de matrizes é tão informativo como dizer que a Nona Sinfonia de Beethoven é uma sequência de vibrações de ar a determinadas frequências. É verdade. Não explica nada que importe.

O que queremos saber, quando perguntamos «porque é que a rede tomou esta decisão?», não é a sequência de operações aritméticas — é a *razão*. O que na entrada provocou a saída. Que características foram determinantes. Que padrão foi reconhecido e esta é uma pergunta que, para as redes neurais modernas, não tem resposta simples nem geral. Não porque alguém tenha decidido não responder, mas porque o processo interno não existe sob a forma de uma narrativa que possa ser traduzida para linguagem humana.

A diferença com o médico é instrutiva precisamente porque o médico também usa conhecimento tácito. Um médico experiente que diagnostica um tumor numa radiografia pode — imperfeitamente, mas de forma inteligível — descrever o que viu: uma opacidade de determinada forma, numa localização específica, com bordos irregulares que não são típicos do tecido normal. O diagnóstico é explicável, pelo menos em traços gerais, em linguagem que outro médico reconhece e pode examinar criticamente. O médico pode estar errado e outro médico pode identificar onde errou.

Uma rede neuronal que diagnostica o mesmo tumor com precisão superior à do

médico não consegue fazer o equivalente. As características que determinaram o diagnóstico existem — são padrões nos pesos das suas camadas profundas —, mas não têm tradução para nenhuma linguagem humana. Não são «bordos irregulares numa localização específica». São geometrias em espaços de centenas de dimensões que não têm nome nem imagem — que, se fossem visualizadas, não corresponderiam a nada que o sistema visual humano reconhecesse como significativo. A opacidade não é parcial. É estrutural.

II. A História da Caixa Negra

A ideia de caixa negra — um sistema cujas entradas e saídas são conhecidas, mas cujo mecanismo interno é opaco — não é nova na ciência. A novidade das redes neuronais não é a opacidade em si, mas a combinação de opacidade profunda com desempenho excepcional, que não tem precedente.

Durante muito tempo, a ciência lidou com caixas negras de forma pragmática e com uma expectativa clara sobre o seu estatuto. Usavam-se modelos empíricos — equações que descreviam as relações entre variáveis sem explicar o mecanismo subjacente — quando o mecanismo ainda não era conhecido. A lei de Boyle, formulada em 1662, descreve a relação inversa entre a pressão e o volume de um gás a temperatura constante, sem explicar porque existe essa relação ao nível molecular. A lei de Ohm, formulada em 1827, descreve a proporcionalidade entre voltagem e corrente sem explicar o que é, microscopicamente, a resistência eléctrica. As leis de Kepler sobre os movimentos planetários eram empiricamente precisas muito antes de Newton as derivar da gravitação universal.

Estas eram caixas negras provisórias — úteis na prática, respeitadas como aproximações válidas, mas reconhecidas como incompletas precisamente porque o mecanismo estava por explicar. A lei de Boyle esperou por Daniel Bernoulli e depois por Maxwell e Boltzmann para receber uma explicação cinética. A lei de Ohm esperou pela física do estado sólido do século XX para ser compreendida microscopicamente. Em cada caso, a abertura da caixa negra representou um salto qualitativo no entendimento — e a expectativa era que todas as caixas negras seriam eventualmente abertas.

Esta expectativa não era uma crença ingénua. Era uma extrapolação do padrão histórico: cada geração de cientistas abria caixas negras que as gerações anteriores tinham tratado como fronteiras do conhecimento. O que não era explicável ontem tornava-se explicável hoje. O futuro pertencia à abertura progressiva, não ao respeito permanente pela opacidade.

As redes neuronais profundas quebram esta expectativa de uma forma que não parece

ser provisória nem corrigível pela acumulação de mais investigação. Não são caixas negras porque ainda não as compreendemos suficientemente bem — são caixas negras por uma razão mais fundamental: o conhecimento que contêm não existe sob uma forma que seja traduzível para a linguagem da explicação causal, independentemente de quanto tempo dedicarmos a tentar. É uma opacidade de tipo diferente.



Existe um campo de investigação — a inteligência artificial explicável, conhecida pelo acrónimo XAI — dedicado precisamente ao problema de tornar as decisões das redes neuronais interpretáveis. Os métodos desenvolvidos são engenhosos e têm utilidade prática real.

O LIME (Local Interpretable Model-agnostic Explanations) treina modelos locais simples — linearmente interpretáveis — que aproximam o comportamento da rede numa vizinhança específica de uma entrada, permitindo identificar quais as características que mais pesam naquela decisão particular. O SHAP (SHapley Additive exPlanations) usa um conceito da teoria dos jogos — os valores de Shapley — para atribuir a cada característica de entrada a sua contribuição marginal para a saída. Os métodos de saliência visual produzem mapas de calor sobre imagens que mostram quais as regiões que mais activaram a resposta da rede. As técnicas de análise de atenção nos modelos de linguagem mostram que palavras do contexto mais influenciaram a geração de cada palavra na saída.

Estes métodos são úteis. Em contextos práticos — auditoria de decisões automáticas, detecção de vieses, depuração de comportamentos inesperados —, fornecem informação valiosa que de outra forma não estaria disponível. Mas são aproximações, não revelações. Identificam correlações entre aspectos da entrada e aspectos da saída, mas não explicam o processo interno pelo qual a rede chegou à sua conclusão — da mesma forma que um mapa de actividade cerebral por ressonância magnética funcional mostra quais as áreas do cérebro que se activam durante uma determinada tarefa sem explicar como essas activações produzem o pensamento.

A analogia com a neuroimagem é instrutiva: passámos décadas a cartografar quais as regiões do cérebro que se activam durante a percepção, a emoção, o raciocínio, a linguagem. Sabemos hoje muito mais sobre a localização funcional do que sabíamos em 1950. Mas a questão de como estas activações produzem a experiência consciente — o problema difícil da consciência — permanece em aberto. A cartografia não é a explicação. A opacidade das redes neuronais é diferente em natureza da opacidade do cérebro, mas o padrão é reconhecível: conhecemos mais sobre o onde do que sobre o como.

III. O que Isto Quebra

Para perceber o que as redes neuronais quebram epistemologicamente, é útil contemplar com mais detalhe o programa positivista que dominou a filosofia da ciência durante grande parte do século XX e que ainda informa, muitas vezes implicitamente, as nossas intuições sobre o que o conhecimento legítimo deve ser.

O positivismo lógico, desenvolvido pelo Círculo de Viena nos anos vinte e trinta do século XX, era uma resposta às pretensões cognitivas da metafísica, da teologia e de outras formas de discurso que faziam afirmações sobre o mundo sem as submeter à verificação empírica. Figuras como Rudolf Carnap, Moritz Schlick e A. J. Ayer argumentavam que o conhecimento científico legítimo tinha de ser verificável empiricamente — que uma proposição que não pudesse ser confirmada ou refutada pela observação não era falsa, era simplesmente sem sentido. Não dizia nada sobre o mundo. Era ruído cognitivo.

Karl Popper, crítico e frequentador do Círculo, mas nunca membro formal, refinaria esta posição com um critério diferente e mais robusto: não a verificabilidade, mas a falsificabilidade. Uma teoria científica tem de fazer previsões que possam, em princípio, ser refutadas pela observação. A teoria da evolução é científica porque prevê que não encontraremos fósseis de mamíferos em estratos do Câmbrio — uma previsão falsificável. A astrologia não é científica porque pode acomodar qualquer resultado: seja o que for que aconteça, há uma interpretação do horóscopo que a torna compatível. O que distingue a ciência da pseudociência não é a capacidade de acumular confirmações, mas a disposição de se expor à possibilidade de estar errada.

Estas posições — verificabilidade e falsificabilidade — partilham um pressuposto que raramente era explicitado porque parecia óbvio demais para precisar de ser dito: que o conhecimento científico tem de ser articulável. Tem de poder ser dito em linguagem precisa, comunicado a outros investigadores, examinado, criticado, revisado. O conhecimento que não pode ser articulado — que existe apenas na mente de um indivíduo, inacessível à verificação pública — não é ciência. É intuição privada, no melhor dos casos. Potencialmente nada.

As redes neuronais profundas possuem conhecimento que não é articulável neste sentido. A sua capacidade de diagnosticar doenças, de traduzir línguas, de reconhecer padrões em dados moleculares ou financeiros ou climáticos é real e demonstrável por qualquer critério de desempenho mensurável. Mas a *forma* desse conhecimento — os pesos, as representações internas, as geometrias em espaços de alta dimensão — não é expressável em linguagem natural nem em nenhuma formalização que humanos possam examinar, criticar e corrigir da forma que a epistemologia tradicional exige. O conhecimento existe. A

articulação não existe e a epistemologia tradicional não tem categoria para este caso.

A pergunta que isto coloca é genuinamente difícil: o que fazemos com este conhecimento? Aceitamo-lo como legítimo porque funciona, mesmo que não o consigamos articular? Tratamo-lo como um novo tipo de conhecimento que exige categorias epistemológicas novas? Ou insistimos que, sem articulação, não é verdadeiro conhecimento — apenas correlação empírica sofisticada que ainda não recebeu a explicação que o tornaria ciência?

Não existe resposta consensual. O que existe é a constatação de que a epistemologia tradicional — construída sobre séculos de ciência com mecanismos articuláveis — não foi concebida para este caso e não o resolve de forma satisfatória. As redes neuronais não são apenas uma nova tecnologia. São um novo tipo de entidade cognitiva que exige que revisitemos algumas das nossas premissas mais fundamentais sobre o conhecimento.



Há uma tensão prática que manifesta esta ruptura filosófica de forma concreta: a tensão entre o desempenho e a confiança — entre o sistema que funciona e o sistema que podemos examinar.

Na medicina, no direito, na engenharia de segurança, nas decisões que afectam vidas humanas, exigimos habitualmente que as decisões sejam justificáveis. Não apenas correctas — justificáveis. O médico pode estar certo no diagnóstico, mas se não consegue explicá-lo de forma que outros médicos possam examinar, o diagnóstico não cumpre os critérios da prática médica responsável. O juiz pode ter chegado à sentença justa, mas se não a fundamenta em razões que possam ser apeladas e revistas, a sentença não cumpre os critérios do Estado de Direito. A justificabilidade não é uma exigência burocrática — é o mecanismo pelo qual os erros são identificados, a responsabilidade é atribuída e o sistema aprende com os seus erros.

Um sistema de IA que diagnostica melhor do que qualquer médico, mas não consegue explicar os seus diagnósticos coloca-nos perante um dilema sem precedentes na história da medicina: confiamos no desempenho e ignoramos a opacidade? Exigimos explicabilidade e aceitamos desempenho inferior? Ou desenvolvemos um novo enquadramento que distingue os contextos em que a opacidade é tolerável dos contextos em que não o é?

Não existe resposta universal e a resposta certa varia claramente com o contexto. Em domínios onde o erro tem consequências limitadas e reversíveis — recomendação de música, filtragem de spam, classificação de fotografias —, a opacidade é tolerável e já a

aceitamos sem grande debate. Mas em contextos de alto risco — diagnóstico de cancro, decisões de liberdade condicional, sistemas de controlo de infraestrutura crítica —, a opacidade cria problemas que não desaparecem por declaração filosófica de que o sistema é muito bom. A bondade do sistema não é suficiente quando o que está em jogo é a responsabilidade, a auditabilidade, a capacidade de identificar e corrigir erros sistemáticos antes que causem danos em larga escala.

O regulamento europeu de IA que entrou em vigor em 2024 reflecte esta intuição: certos sistemas de alto risco devem ser explicáveis, transparentes, auditáveis. É uma resposta política a uma questão filosófica — e como todas as respostas políticas a questões filosóficas, é imperfeita. Mas é um reconhecimento de que a tensão entre desempenho e confiança é real e tem de ser gerida, não ignorada.

IV. Uma Ruptura, não um Defeito

É tentador — e errado — tratar a opacidade das redes neuronais como um problema temporário à espera de solução técnica. Como se, com mais investigação e métodos melhores de interpretabilidade, a caixa negra se abrisse progressivamente e o mecanismo se tornasse tão transparente como o de uma equação diferencial ou de um programa lógico.

Esta expectativa confunde dois tipos diferentes de opacidade que é essencial distinguir. A opacidade provisória é a do mecanismo ainda não compreendido que a investigação eventualmente ilumina. É o tipo de opacidade que a lei de Boyle tinha antes da teoria cinética dos gases, ou que a genética mendeliana tinha antes da descoberta da estrutura do ADN. É real e resistente, mas é em princípio dissolvível pela investigação adequada — porque o mecanismo existe sob uma forma que é, em princípio, redutível à linguagem da explicação científica.

A opacidade estrutural é diferente em natureza. É a opacidade do conhecimento que existe numa forma que é genuinamente irreduzível à articulação verbal ou formal — não por falta de investigação, mas porque o tipo de representação em que o conhecimento existe não tem equivalente na linguagem que os humanos usam para articular conhecimento. O conhecimento de uma rede neuronal profunda é estruturalmente opaco porque existe sob a forma de padrões distribuídos em espaços de altíssima dimensão, sem localização, sem estrutura narrativa, sem os tipos de regularidade local que a linguagem consegue capturar.

A distinção é como a diferença entre a dificuldade de traduzir um poema de uma língua para outra — difícil, mas em princípio possível com esforço suficiente e criatividade — e a impossibilidade de traduzir uma experiência sensorial directa para linguagem proposicional. Posso tentar descrever o sabor do vinho que estou a beber e a descrição pode ser boa ou

má, mais ou menos evocativa. Mas nenhuma descrição, por mais elaborada que seja, é o sabor — e alguém que nunca provou vinho não terá, através da descrição, a mesma experiência que eu tenho. A experiência não tem a forma de uma proposição.

O conhecimento das redes neuronais é do segundo tipo. Isto não é uma metáfora: é uma afirmação técnica sobre a estrutura das representações internas que as redes desenvolvem. As pesquisas em interpretabilidade mecânica — um campo que procura compreender o que exactamente as redes neuronais aprendem ao nível dos seus componentes — têm mostrado que as redes desenvolvem estruturas internas sofisticadas e, por vezes, surpreendentes. Mas estas estruturas existem em espaços de alta dimensão que não têm correspondência nas categorias com que os humanos organizam o conhecimento. Visualizá-las é possível com técnicas de redução de dimensionalidade; compreendê-las no sentido em que compreendemos um mecanismo — sendo capazes de prever o que aconteceria se mudássemos esta componente ou aquela — é uma questão diferente e muito mais difícil.



Há algo de paradoxalmente familiar nesta situação que ajuda a perceber porque é que a opacidade das redes, embora filosoficamente perturbadora, não é intelectualmente paralisante.

Os humanos vivem com conhecimento não articulável desde sempre — desde muito antes de existirem redes neuronais para tornar o fenómeno saliente. O pianista que não consegue explicar como produz o fraseado certo. O cirurgião que percebe pela resistência dos tecidos que está no plano correcto. O professor experiente que sabe, no primeiro dia de aulas, quais os estudantes que vão ter dificuldades, sem ser capaz de articular os sinais que o informaram. Todos eles possuem conhecimento que funciona — que é real, eficaz e fiável no seu domínio — e que não é completamente articulável.

Quando este conhecimento estava confinado à mente humana, podia ser tratado como uma curiosidade filosófica ou como uma limitação contingente: talvez, com mais introspecção, com melhores técnicas de análise cognitiva, com mais neuroimagem, o conhecimento pudesse ser articulado. A opacidade da mente humana era real, mas parecia, em princípio, dissolúvel — era uma questão de instrumentos insuficientes, não de impossibilidade de princípio.

Quando esse mesmo tipo de conhecimento existe numa máquina construída por nós, com matemática que compreendemos completamente ao nível das operações elementares, a ilusão de que a articulabilidade é apenas uma questão de esforço insuficiente torna-se

muito mais difícil de manter. A rede é feita de multiplicações de matrizes que entendemos perfeitamente. O resultado dessas multiplicações é um conhecimento que não conseguimos articular. A opacidade não vem de instrumentos insuficientes — vem da natureza do processo.

As redes neuronais não criaram o problema do conhecimento tácito. Tornaram-no impossível de ignorar como contingência removível — forçaram-nos a reconhecê-lo como uma característica estrutural de certos tipos de processos cognitivos, biológicos ou artificiais.

E ao tornarem-no impossível de ignorar, forçaram-nos a olhar de forma nova para a inteligência — a nossa, a delas — e para a relação entre saber e compreender que dá nome a este livro. Porque se o conhecimento pode existir sem articulação, então a questão de o que é saber — de o que é ter conhecimento — precisa de ser recolocada. É essa recolocação que os capítulos da Parte III vão empreender, virando o espelho das redes neuronais directamente para a mente humana.



PARTE III

*O Espelho Humano: Intuição e Conhecimento
Inexprimível*

CAPÍTULO 6

A Intuição das Máquinas e a Cognição Humana

"A intuição não é um dom misterioso. É o reconhecimento de padrões."

Herbert Simon, *Models of My Life*

I. O que Queremos Dizer quando Dizemos Intuição

A palavra «intuição» carrega um peso filosófico incómodo que começa na sua própria história etimológica. Do latim *intueri* — olhar directamente, contemplar —, passou a designar um modo de conhecimento que acede à verdade sem a mediação do raciocínio discursivo. Para os filósofos medievais, era o modo pelo qual Deus conhecia tudo de uma só vez, em contraste com o conhecimento humano que procede passo a passo. Para os racionalistas modernos, era a capacidade da razão de apreender verdades evidentes por si mesmas — os axiomas a partir dos quais o raciocínio dedutivo parte. Para os românticos, era o dom do génio: a faísca de inspiração que não pode ser ensinada nem aprendida.

Esta herança tornava a intuição filosoficamente respeitável, mas cientificamente inútil. Se a intuição é por definição inacessível à análise, então não pode ser estudada, ensinada, melhorada, ou sequer compreendida. É um mistério que se invoca quando a explicação falha — o equivalente cognitivo do «é magia».

A psicologia cognitiva das últimas décadas propôs uma leitura radicalmente diferente. A intuição não é magia. É reconhecimento de padrões ultra-rápido, produzido por um sistema cognitivo que processou grandes quantidades de experiência e aprendeu a identificar configurações relevantes sem necessidade de raciocínio explícito. É um produto da aprendizagem, não um substituto para ela. É o resultado de um treino longo e exigente que se tornou tão fluido que deixou de ser visível.

Esta redefinição não diminui a intuição — revela a sua verdadeira natureza e, nesse sentido, torna-a mais impressionante, não menos. O diagnóstico instantâneo do médico experiente não é menos notável por ser o produto de milhares de horas de prática; é mais notável, porque revela que o sistema cognitivo humano pode comprimir décadas de experiência em padrões de reconhecimento que operam à velocidade da percepção.

Herbert Simon, prémio Nobel de Economia e um dos fundadores da ciência cognitiva, chegou a esta conclusão através de uma série de estudos sobre jogadores de xadrez de topo

que se tornaram clássicos da literatura. Simon e o seu colaborador William Chase desenvolveram o paradigma da recordação de posições de xadrez: mostravam aos participantes — grandes mestres, jogadores intermédios, principiantes — posições de tabuleiro durante alguns segundos e pediam-lhes depois que as reproduzissem. Os grandes mestres recordavam quase todas as peças com exactidão. Os principiantes recordavam apenas quatro ou cinco.

A conclusão óbvia — que os grandes mestres têm memória superior — foi refutada por um controlo elegante: quando as peças eram colocadas de forma aleatória, sem seguir nenhuma lógica de jogo real, a vantagem dos grandes mestres desaparecia completamente. Recordavam as posições aleatórias tão mal quanto os principiantes. O que os grandes mestres tinham não era memória superior — era um vocabulário de padrões. Reconheciam configurações com significado no xadrez como palavras conhecidas num texto; as configurações aleatórias eram como texto numa língua desconhecida.

Simon estimou que um grande mestre acumula, ao longo da sua carreira, entre cinquenta mil e cem mil padrões de posição reconhecíveis. Quando olha para um tabuleiro e «vê» imediatamente a jogada certa — sem calcular explicitamente as variantes —, não está a ter uma visão mística. Está a activar um padrão armazenado que a experiência lhe ensinou a associar a certas jogadas eficazes. A intuição do grande mestre é, na formulação de Simon, *estatística implícita*: o resultado de uma longa exposição a dados — neste caso, posições de xadrez e os seus resultados — que o sistema cognitivo processou e condensou em heurísticas de reconhecimento que operam abaixo do limiar da consciência explícita.



Esta definição de intuição é importante por uma razão que vai além da psicologia cognitiva: ela torna a comparação com as redes neuronais não apenas metafórica, mas conceptualmente precisa — e iluminadora em ambas as direcções.

Uma rede neuronal profunda treinada para reconhecer tumores em tomografias computadorizadas é, funcionalmente, um sistema que processou grandes quantidades de imagens médicas e aprendeu a identificar configurações relevantes — padrões que se associam a determinados diagnósticos — sem que essas configurações tenham sido explicitamente definidas por ninguém. O processo pelo qual chega às suas conclusões é rápido, automático, não mediado por raciocínio explícito passo a passo e as configurações que aprendeu a reconhecer não são articuláveis em linguagem natural.

Se aplicarmos a definição de Simon à rede neuronal — se perguntarmos se ela acumulou, através da exposição a exemplos, padrões de reconhecimento que opera sem

raciocínio explícito e que não podem ser completamente verbalizados —, a resposta é afirmativa. A descrição encaixa. A rede tem o que a psicologia cognitiva chama de intuição especializada, pelo menos no sentido operacional e funcional do termo.

Mas aqui é necessária uma precisão que este capítulo vai manter ao longo de toda a sua extensão. Esta convergência funcional não significa que a rede neuronal e o grande mestre de xadrez estejam a fazer a mesma coisa nos mesmos sentidos que importam para a filosofia da mente. A rede não «sente» o tabuleiro. Não tem a experiência subjectiva de reconhecer um padrão. Não tem a história pessoal de derrotas e vitórias que deu significado aos padrões. A analogia é uma ferramenta conceptual — ilumina algo real sobre a estrutura do reconhecimento de padrões —, mas tem limites que precisam de ser respeitados. Enunciaremos esses limites à medida que avançarmos.

II. Os Dois Sistemas e a Máquina

Daniel Kahneman, no seu influente «Thinking, Fast and Slow», sistematizou uma distinção que a psicologia cognitiva vinha a desenvolver há décadas e que organiza de forma elegante uma grande quantidade de evidência experimental: entre dois modos de processamento mental a que chamou Sistema 1 e Sistema 2. A sistematização de Kahneman não é a única possível nem é universalmente aceite pelos seus colegas, mas captura algo real sobre a arquitectura da cognição humana de uma forma suficientemente precisa para ser útil.

O Sistema 1 é rápido, automático, associativo, não deliberativo. Opera em paralelo, sem esforço consciente, produzindo respostas imediatas com base em padrões reconhecidos. É o sistema que lê as palavras desta página sem deliberação consciente sobre cada letra, que reconhece a emoção num rosto antes de poder nomeá-la, que sente imediatamente que « $2 + 2$ » é «4» sem calcular. É o sistema que detecta que algo está errado no tom de uma conversa antes de conseguir articular o quê. É o sistema que guia a condução automática num percurso familiar enquanto a mente consciente pensa noutras coisas.

O Sistema 1 é também o sistema que produz preconceitos, heurísticas e erros sistemáticos que Kahneman e Amos Tversky documentaram extensivamente ao longo de décadas de investigação. A heurística da disponibilidade — a tendência para avaliar a probabilidade de um evento com base na facilidade com que exemplos ocorrem à memória — é um produto do Sistema 1. O viés de ancoragem — a influência desproporcionada do primeiro número apresentado numa negociação — é um produto do Sistema 1. A aversão à perda — a tendência para sentir as perdas com maior intensidade do que os ganhos equivalentes — é um produto do Sistema 1. O sistema é poderoso e eficiente, mas não é

neutro.

O Sistema 2 é lento, deliberativo, sequencial, esforçado. É o sistema que calcula « 17×24 », que verifica um argumento lógico passo a passo, que reconsidera uma decisão impulsiva à luz de consequências futuras, que pondera deliberadamente as opções antes de uma escolha importante. Requer atenção consciente e consome recursos cognitivos mensuráveis — o que é demonstrado pelo facto de que a resolução de problemas matemáticos difíceis dilata as pupilas e aumenta a frequência cardíaca, sinais de activação fisiológica que o reconhecimento de rostos não produz.

Kahneman argumenta que a maior parte do pensamento quotidiano é feito pelo Sistema 1 — que o Sistema 2 é chamado apenas quando o Sistema 1 encontra uma situação que não reconhece, ou quando é explicitamente convocado para verificar as conclusões do Sistema 1, ou quando a tarefa requer o tipo de atenção sequencial que o Sistema 1 não consegue prestar. A mente humana é, na sua operação normal, muito menos deliberativa do que gostamos de acreditar. A ilusão de que pensamos de forma consciente e racional na maioria das situações é ela própria um produto do Sistema 1, que constrói narrativas post hoc para as decisões que já tomou.

As redes neuronais profundas são, funcionalmente, quase inteiramente Sistema 1. Processam entradas e produzem saídas de forma rápida, paralela, não deliberativa. Não existe um processo de verificação sequencial, nenhuma consulta de regras explícitas, nenhuma deliberação passo a passo que pondere alternativas e escolha a melhor. Há apenas o fluxo de activações através das camadas, produzindo uma saída que é o produto de tudo o que a rede aprendeu. Da mesma forma que o Sistema 1 humano produz uma resposta que é o produto de toda a experiência acumulada.



Esta correspondência funcional entre as redes neuronais e o Sistema 1 humano tem implicações tanto para a compreensão das redes como para a compreensão do Sistema 1. Do lado das redes: sugere que o seu modo de funcionamento não é uma aberração técnica, mas corresponde ao modo de funcionamento de uma parte substancial da cognição natural. Do lado do Sistema 1: sugere que o que o Sistema 1 faz é mais parecido com computação do que a tradição filosófica admitia.

O Sistema 1 humano é extraordinariamente poderoso nos domínios onde acumulou experiência de alta qualidade. O médico que reconhece o diagnóstico ao entrar no quarto, o sommelier que identifica a colheita pelo aroma sem ver o rótulo, o músico que detecta a nota errada no ensaio antes de poder nomear qual é, o bombeiro que sente que o edifício

está prestes a colapsar antes de ver os sinais físicos explícitos — todos eles estão a usar um Sistema 1 treinado por milhares de horas de exposição ao domínio, em condições que forneceram feedback suficiente para calibrar os padrões aprendidos.

Mas o Sistema 1 falha sistematicamente fora dos domínios de experiência e falha de formas que revelam a natureza do seu funcionamento: produz respostas baseadas em similaridade superficial em vez de estrutura profunda; é sensível ao enquadramento e à ordem de apresentação da informação, mesmo quando esta não tem relevância lógica para a decisão; é impermeável à lógica quando o padrão activado é suficientemente forte; e constrói confiança independentemente da fiabilidade, o que produz especialistas que se sentem certos mesmo quando estão errados.

As redes neuronais profundas exibem falhas que são funcionalmente análogas. São extraordinariamente competentes nos domínios e nas distribuições de dados em que foram treinadas. Mas falham de formas que revelam exactamente a mesma arquitectura epistemológica: são vulneráveis a «exemplos adversariais» — perturbações subtis da entrada, imperceptíveis para humanos, que produzem classificações completamente erradas com alta confiança; generalizam mal quando a distribuição dos dados de teste difere significativamente da distribuição de treino; e exibem a mesma dissociação entre confiança e fiabilidade que o Sistema 1 humano. A simetria das falhas é uma das evidências mais fortes de que a analogia captura algo real.

III. A Questão da Fiabilidade

Kahneman é enfático num ponto que é frequentemente esquecido nas discussões entusiastas sobre intuição especializada: a intuição é fiável apenas quando se formou em condições específicas que ele descreve com dois critérios relacionados.

O primeiro critério é a *validade do ambiente*: a intuição é fiável quando o domínio em que se formou tem regularidades estáveis e aprendíveis — quando existe uma relação previsível entre os sinais que o sistema percebe e os resultados que se seguem. O xadrez é um ambiente de alta validade: as configurações do tabuleiro têm consequências determinísticas, as relações entre posição e resultado são estáveis e o feedback sobre a qualidade das jogadas é imediato e inequívoco. A medicina clínica, em muitos domínios, também: os padrões de imagem radiológica correspondem de forma relativamente estável a condições anatómicas e os diagnósticos são validados por patologia, cirurgia ou evolução clínica.

O segundo critério é a *quantidade e qualidade do feedback*: a intuição só se calibra correctamente quando existe feedback frequente, rápido e claro sobre a adequação das respostas. O jogador de xadrez recebe feedback imediato — ganha ou perde e pode analisar

as suas decisões. O médico recebe feedback, embora por vezes com atraso. O piloto recebe feedback através dos instrumentos e da resposta do avião. Em domínios onde o feedback é lento, ruidoso, ou sistematicamente enviesado — onde os erros não são visíveis ou onde as consequências chegam tarde demais para serem associadas às decisões que as causaram —, a intuição não se calibra, mesmo com muita experiência.

Kahneman cita as previsões financeiras de longo prazo como exemplo paradigmático de intuição em ambiente de baixa validade. Os analistas financeiros desenvolvem uma confiança subjectiva forte nas suas previsões — sentem que sabem o que vai acontecer ao mercado. Os estudos mostram que o seu desempenho é estatisticamente indistinguível do acaso para horizontes temporais longos. A experiência e a confiança existem; a fiabilidade não. O que se desenvolveu com a experiência foi a *sensação* de saber, não o saber.

Esta distinção entre a confiança subjectiva e a fiabilidade objectiva é um dos contributos mais importantes e mais perturbadores da psicologia cognitiva. Perturbador porque sugere que a experiência, por si só, não garante competência — que é possível ter décadas de prática e desenvolver padrões de reconhecimento que são sistemática e confiantemente errados, se as condições de aprendizagem não forneceram feedback calibrado.

A implicação para as redes neuronais é directa e importante. Uma rede treinada num ambiente de alta validade — onde os dados de treino correspondem de forma estável e representativa à realidade que se pretende modelar — desenvolve capacidades genuinamente fiáveis que generalizam para novos dados da mesma distribuição. Uma rede treinada em dados enviesados, não representativos, ou que reflectem correlações espúrias do conjunto de treino, desenvolve padrões que parecem robustos dentro do domínio de treino, mas que falham sistematicamente quando implantados no mundo real.

Os casos históricos de fracasso de sistemas de IA em implantação real são, quase invariavelmente, casos de treino em ambientes de baixa validade. O sistema de reconhecimento facial que funciona bem em testes laboratoriais e falha sistematicamente para rostos de certas etnias sub-representadas nos dados de treino. O modelo de diagnóstico médico treinado num hospital universitário de elite que generaliza mal para populações com diferentes perfis socioeconómicos. O sistema de crédito que aprende a discriminar não porque a discriminação foi programada, mas porque as correlações nos dados históricos reflectiam padrões de discriminação passados.

Estes fracassos não são acidentes técnicos — são as consequências previsíveis de treinar sistemas de reconhecimento de padrões em ambientes de baixa validade. São o equivalente exacto, na rede neuronal, do analista financeiro que desenvolveu confiança sem

fiabilidade: o sistema aprendeu padrões reais nos dados de treino, mas esses padrões não correspondem às estruturas causais do mundo real que se pretendia modelar.

IV. O que a Máquina Demonstra

Chegamos ao ponto central deste capítulo, que é também o ponto central de toda a Parte III. Vale a pena enunciá-lo com precisão, porque é fácil dizer demasiado ou de menos.

Não estamos a afirmar que as redes neuronais têm intuição no sentido pleno em que os humanos têm intuição — com experiência subjectiva do reconhecimento, com a história pessoal que dá significado aos padrões, com o horizonte existencial que torna o saber mais do que processamento de informação. Não estamos a afirmar que o processo computacional de uma rede neuronal é equivalente ao processo cognitivo de um grande mestre de xadrez. Não estamos a colapsar a distinção entre o sistema de silício e o sistema de carne.

Estamos a afirmar algo mais preciso e, creio, mais interessante: que as redes neuronais constituem uma prova de conceito epistemológica. Demonstram empiricamente que o reconhecimento de padrões por indução estatística — sem regras explícitas, sem princípios inatos, sem raciocínio deliberativo, apenas por exposição a exemplos — é suficiente para produzir capacidades que, quando exibidas por humanos, classificamos como inteligência, perícia, intuição especializada.

Esta prova de conceito tem implicações filosóficas que vão além da engenharia. Se um sistema puramente indutivo — que não tem biologia, não tem corpo, não tem história evolutiva, não tem nada além de uma arquitectura matemática e dados de treino — consegue igualar ou superar o desempenho humano em domínios que julgávamos requerer compreensão profunda, então a nossa explicação habitual do que torna o desempenho humano possível nesses domínios precisa de ser revisitada.

O exemplo mais perturbador é o diagnóstico médico. A explicação tradicional do que torna um bom radiologista é que ele compreende a anatomia normal e patológica — tem um modelo conceptual do que deve parecer um tecido saudável, do que os diferentes tipos de lesão produzem em imagem, de como a doença progride e se manifesta. Esta compreensão permite-lhe reconhecer o que é significativo numa imagem, distinguir variantes normais de anomalias relevantes, integrar a informação de imagem com o contexto clínico.

As redes neuronais que igualam ou superam o diagnóstico radiológico humano não têm nada disto. Não têm modelo conceptual da anatomia. Não têm compreensão da progressão da doença. Não têm contexto clínico que orientem. Têm padrões estatísticos

extraídos de centenas de milhares de imagens anotadas e isso — apenas isso — é suficiente para produzir o desempenho.

O que isto sugere não é que o radiólogo não precisa de compreender anatomia — a compreensão tem funções importantes que vão além do diagnóstico por padrões: guia a atenção para áreas relevantes, informa a comunicação com outros clínicos, orienta as decisões terapêuticas, permite raciocinar sobre casos genuinamente atípicos. O que sugere é que a parte operativamente mais poderosa do diagnóstico — a identificação inicial da anomalia — é produzida por um processo de reconhecimento de padrões que tem mais em comum com o funcionamento de uma rede neuronal do que com o raciocínio hipotético-dedutivo que os manuais de metodologia clínica descrevem. A *compreensão* que o radiólogo acredita ter pode ser, em parte substancial, uma narrativa post hoc que o Sistema 2 constrói para racionalizar o que o Sistema 1 já decidiu.



Há uma ressalva final que este capítulo deve enunciar com cuidado — não como codicilo burocrático, mas como delimitação genuinamente importante do argumento.

A diferença de eficiência de amostra entre humanos e redes neurais é real e significativa e não foi resolvida pela escala. Uma criança aprende o conceito de «cadeira» vendo meia dúzia de cadeiras em diferentes contextos, ouvindo a palavra usada por adultos, sentando-se nelas, caindo de algumas. Uma rede neuronal precisa de dezenas de milhares de imagens rotuladas para aprender a categoria com fiabilidade comparável. Esta assimetria não é trivial: indica que a mente humana tem estruturas inatas que orientam a aprendizagem de forma que os sistemas puramente indutivos não têm — predisposições para aprender certas categorias (objectos, faces, agentes, relações causais) de forma muito mais eficiente do que a indução pura a partir de dados brutos permitiria.

A diferença de corporeidade é igualmente real. A intuição diagnóstica do médico não se formou apenas em imagens: formou-se em contacto com doentes reais, com a textura da pele, o cheiro característico de certas condições, a postura do corpo, a expressão facial que acompanha a dor crónica versus a aguda. É uma intuição incorporada — que existe no corpo do médico tanto quanto na sua mente — num sentido que nenhuma rede treinada em imagens digitais pode reivindicar. O corpo não é apenas um veículo para a mente; é uma das fontes do conhecimento e em muitos domínios a mais importante.

Estas diferenças são importantes e devem ser levadas a sério. Mas não invalidam o que a comparação revela: que o princípio do reconhecimento de padrões por indução estatística — o princípio que Hume descreveu como a base do conhecimento empírico e que o

paradigma racionalista subestimou durante séculos — é suficientemente poderoso para produzir comportamento inteligente a uma escala que ninguém antecipava e que a fronteira entre o que requer «verdadeira compreensão» e o que pode ser alcançado por reconhecimento de padrões sofisticado é muito menos nítida do que o racionalismo cartesiano nos levou a acreditar.

É desta fronteira esbatida — e do que ela nos revela sobre a nossa própria mente — que o próximo capítulo trata directamente.



CAPÍTULO 7

O que a Máquina nos Diz sobre Nós Próprios

"A mente não é um vaso a encher, mas um lume a acender."

Plutarco, Da Escuta

I. O Espelho Inesperado

Há uma ironia que atravessa toda a história da inteligência artificial e que raramente é contemplada na sua plena extensão: ao tentar construir máquinas inteligentes, aprendemos mais sobre a inteligência humana do que em séculos de filosofia e introspecção e aprendemos mais, paradoxalmente, pelos fracassos do que pelos sucessos.

Não porque as máquinas se tenham revelado iguais aos humanos. Pelo contrário: as formas específicas como falharam, os problemas particulares que não conseguiram resolver, as capacidades que demoraram décadas a aproximar — tudo isso funcionou como um espelho que reflectiu, em negativo, a forma da inteligência natural. Cada fracasso da IA era uma lição sobre o que a inteligência humana faz que não é óbvio nem trivial — que estava tão completamente incorporado na nossa forma de estar no mundo que nunca nos tínhamos dado ao trabalho de o examinar.

O fracasso da tradução automática baseada em regras mostrou que a compreensão linguística não é a aplicação de gramáticas — é a imersão num mundo partilhado de referências, intenções e contextos que nenhuma gramática captura. Falantes nativos desambigam instantaneamente frases que os analisadores sintácticos mais sofisticados tratavam como profundamente ambíguas, não porque conheçam mais regras gramaticais, mas porque partilham um mundo de experiências e convenções que ancora a interpretação.

O fracasso do reconhecimento de objectos por visão computacional clássica mostrou que ver não é detectar arestas e calcular geometrias — é reconhecer configurações num espaço de alta dimensão que a experiência ensinou a categorizar. Reconhecemos a nossa mãe num rosto parcialmente tapado, numa fotografia desfocada, de um ângulo nunca visto antes. Os sistemas de visão que tentavam codificar explicitamente as características relevantes falhavam com uma consistência que revelava quão profundamente diferente é o processo humano.

O fracasso dos sistemas especialistas em medicina mostrou que o diagnóstico não é a aplicação de árvores de decisão — é o reconhecimento holístico de padrões que integra

dimensões que nenhum manual enumera completamente. O «aspecto geral» do doente que os médicos descrevem — a percepção imediata de que algo está errado antes de identificar o quê — resiste completamente à formalização em regras, porque não existe como conjunto de proposições: existe como percepção directa.

Em cada caso, o fracasso da abordagem explícita revelou que a competência humana correspondente era, em grande medida, tácita. Que o «segredo» não estava nas regras. Estava na experiência acumulada que tinha transformado os padrões em reconhecimento automático, tão fluido e rápido que a própria pessoa competente não o percebia como um processo.

O sucesso das redes neuronais não desfaz este espelho — aprofunda-o e torna-o mais perturbador. Porque ao mostrar que o reconhecimento de padrões por indução é suficiente para igualar o desempenho humano, as redes neuronais levantam uma pergunta ainda mais desconfortável: se uma máquina consegue fazer o que o especialista humano faz apenas com indução estatística a partir de dados, quanto da «compreensão» que o especialista acredita ter é, afinal, esse mesmo processo de indução? Quanto do que tomávamos por pensamento profundo é, na sua parte operativamente mais poderosa, reconhecimento de padrões?



Esta pergunta não é retórica. Tem consequências reais e mensuráveis para a forma como entendemos a nossa própria cognição e a evidência experimental que se acumulou nas últimas décadas torna-a difícil de ignorar.

Considere um radiólogo com vinte anos de experiência. Ele olha para uma tomografia e produz um diagnóstico. Perguntado sobre o seu raciocínio, descreve um processo que parece profundamente conceptual: identificou esta característica na imagem, que é consistente com aquela condição, considerou as alternativas, pesou as probabilidades clínicas, integrou a informação com o contexto do doente. A narrativa é coerente, articulada, marcada pela linguagem da compreensão analítica e do raciocínio hipotético-dedutivo.

Mas existe uma quantidade crescente de evidência experimental de que esta narrativa é, em grande parte, uma reconstrução post hoc — uma história que o Sistema 2 conta sobre uma decisão que o Sistema 1 já tomou. Os estudos de rastreio ocular de radiologistas experientes mostram que fixam a região patológica nos primeiros segundos de observação, antes de poderem nomear conscientemente o que vêem — a atenção é atraída para a anomalia antes de a anomalia ser identificada explicitamente. Os estudos de tempo de

resposta mostram que a velocidade do diagnóstico de especialistas é incompatível com o raciocínio deliberativo que descrevem: o processo que narram seria muito mais lento do que o processo que efectivamente ocorre.

Isto não significa que a justificação seja inútil — longe disso. O processo de verificação explícita que o Sistema 2 realiza serve funções cruciais: detecta erros do Sistema 1, comunica o raciocínio a outros clínicos, documenta para fins legais e de auditoria e proporciona o escrutínio que permite ao sistema médico aprender com os seus erros. Mas sugere que a parte operativamente mais poderosa do diagnóstico — a identificação inicial da anomalia — é produzida por um processo que tem muito mais em comum com o reconhecimento de padrões de uma rede neuronal do que com o raciocínio hipotético-dedutivo que os manuais de metodologia clínica descrevem como o ideal do pensamento médico.

II. Hume Tinha Razão — e as Redes Mostram Porquê

David Hume, escrevendo no século XVIII numa Edimburgo que ficaria conhecida como a Atenas do Norte, propôs uma das teorias da mente mais radicais e mais férteis da história da filosofia ocidental. O projecto de Hume era nada menos do que uma ciência da natureza humana — uma descrição rigorosa do funcionamento real da mente, em contraste com as especulações metafísicas que dominavam a filosofia da sua época.

Contra Descartes, que acreditava que a mente possuía ideias inatas — verdades que o pensamento encontra em si mesmo, independentemente de qualquer experiência — e que o conhecimento certo podia ser derivado pela razão pura a partir destas verdades, Hume argumentou de forma sistemática que toda a actividade mental tem a sua origem em impressões. Por impressões Hume entendia experiências sensoriais concretas: o vermelho que vejo, o frio que sinto, o barulho que ouço. As ideias — os conceitos abstractos com que o pensamento opera — são cópias mais fracas e menos vívidas destas impressões originais.

Para Hume, não existe conhecimento a priori no sentido forte que Descartes reclamava. Não existe uma razão pura que derive verdades necessárias sobre o mundo empírico. Existe apenas a experiência — o fluxo contínuo de impressões — e o hábito que nos leva a associar certas impressões com certas outras, a esperar que o que sempre aconteceu depois de A continue a acontecer depois de A. O conhecimento é, na sua essência, associação habitual de padrões.

A famosa análise humiana da causalidade é o exemplo mais provocador desta posição. Quando vemos a bola de bilhar A colidir com a bola B e B mover-se, não observamos a *causa*. Observamos a sequência. A ideia de que A *causa* B — de que existe uma ligação

necessária entre os dois eventos — não é uma observação. É uma inferência produzida pelo hábito de ver sequências repetidas. A causalidade não está no mundo; está na mente que aprendeu a associar padrões. Hume não estava a negar que as bolas de bilhar colidem e que a colisão produz movimento — estava a dizer que a *necessidade* desta relação, a ideia de que *tinha* de acontecer assim, é uma projecção da nossa experiência repetida sobre o mundo, não uma leitura de algo que está objectivamente lá.

Esta posição era escandalosa no século XVIII — Kant diria que Hume o despertou do seu «sono dogmático», forçando-o a repensar os fundamentos do conhecimento — e continua a ser filosoficamente controversa. Mas as redes neuronais oferecem uma perspectiva nova sobre ela que Hume não podia antecipar: não como argumento filosófico, mas como demonstração empírica que torna o argumento de Hume muito mais difícil de refutar.

Uma rede neuronal não tem acesso a relações causais no sentido profundo. Aprende correlações — aprende que certas configurações de entrada estão associadas a certas saídas, repetidamente, nos dados de treino. O seu «conhecimento» do mundo é, exactamente no sentido de Hume, um conjunto de associações habituais derivadas da exposição repetida a padrões. Não há causalidade no processo; há correlação estatística suficientemente consistente para funcionar como base de previsão e de acção.

E no entanto funciona. Funciona extraordinariamente bem em domínios muito variados. A rede que prevê a estrutura de proteínas, a rede que traduz entre línguas, a rede que reconhece tumores — todas elas operam por associação estatística de padrões, sem acesso a relações causais profundas e todas elas produzem resultados que a ciência, a comunicação e a medicina precisam. O que a rede neuronal demonstra — de forma que nenhum argumento filosófico poderia demonstrar tão concretamente — é que o conhecimento associativo, por si só, é suficiente para produzir comportamento que classificamos como inteligente em domínios que exigem décadas de estudo humano.

Não estamos a dizer que Hume estava completamente certo. A psicologia cognitiva moderna mostra que a mente humana tem estruturas inatas que guiam a aprendizagem de formas que o empirismo radical de Hume não antecipava — a predisposição para aprender linguagem em janelas temporais específicas, para reconhecer agentes causais e atribuir-lhes intenções, para construir representações de objectos persistentes que permanecem mesmo quando não são vistos. A mente não é a Tabula Rasa em branco que Locke descreveu e os bebés não chegam ao mundo sem nenhuma estrutura cognitiva pré-formada.

Mas estamos a dizer que Hume capturou algo essencial sobre o mecanismo pelo qual o conhecimento empírico funciona — que a aprendizagem por associação de padrões é

mais poderosa do que o racionalismo admitia e que as redes neuronais tornaram esse algo visível de uma forma que a filosofia por si só, sem o auxílio de máquinas que aprendem, não podia alcançar.

III. O Racionalismo Revisitado

É justo, neste ponto, dar voz à posição contrária — não como espantalho a abater para facilitar o argumento, mas como objecção séria que o leitor rigoroso já está a formular e que merece uma resposta directa.

O racionalista pode argumentar com justeza: muito bem, as redes neuronais são impressionantes em domínios de reconhecimento de padrões. Mas há domínios onde o reconhecimento de padrões por indução claramente não chega — domínios que são precisamente os que a tradição filosófica considerava o coração da racionalidade humana.

A matemática pura é o exemplo mais óbvio. Demonstrar um teorema novo não é reconhecer um padrão visto antes nos dados de treino — é construir um argumento dedutivo que tem de ser válido em todos os casos possíveis, incluindo os infinitos casos que nunca ocorreram nos dados de treino. A prova de um teorema é uma construção que estabelece a verdade necessária de uma proposição pela sua relação com axiomas, não pela sua semelhança com exemplos passados. É exactamente o tipo de operação que o modelo associativo de Hume não consegue capturar completamente.

O raciocínio contrafactual é outro exemplo perturbador. A capacidade de raciocinar sobre o que teria acontecido se as coisas tivessem sido diferentes — «se a penicilina não tivesse sido descoberta, a Segunda Guerra Mundial teria tido mais mortes por infecção» — é essencial ao planeamento, à ética, à ciência e ao direito. Exige raciocinar sobre mundos que não existem e portanto sobre situações para as quais não há dados de treino. As redes neuronais actuais mostram limitações significativas precisamente neste domínio.

A objecção é pertinente e a evidência disponível suporta-a em parte. As redes neuronais actuais têm dificuldades reais com raciocínio abstracto puro, com generalização consistente fora da distribuição de treino, com a construção de argumentos dedutivos que se mantenham válidos independentemente dos exemplos vistos. Há algo que o racionalismo chama «razão» — uma capacidade de raciocinar a partir de princípios gerais para casos particulares, de forma que seja válida independentemente da experiência passada — que não se reduz facilmente ao reconhecimento de padrões estatístico.

A admissão desta objecção é genuína. Não a estamos a fazer por condescendência retórica. Os defensores mais rigorosos das redes neuronais reconhecem estas limitações como problemas técnicos abertos, não como questões resolvidas. O debate sobre se as

redes neuronais actuais «raciocinam» no sentido pleno, ou apenas aparentam raciocinar por combinação sofisticada de padrões linguísticos, é um debate legítimo com posições sérias em ambos os lados.

Mas — e este «mas» é importante — a admissão da objecção não repõe o racionalismo cartesiano no seu trono. O que mostra é que a inteligência humana plena requer provavelmente ambos os modos: o reconhecimento de padrões que as redes neuronais simulam com extraordinária competência e algo mais — talvez a capacidade de construir modelos causais explícitos, talvez o raciocínio simbólico combinado com o padrão-reconhecimento, talvez formas de abstracção que ainda não sabemos reproduzir artificialmente.

O que as redes neuronais desafiaram de forma irreversível — e este desafio não é afectado pelas suas limitações actuais — é a ideia de que a inteligência é essencialmente o segundo modo. Que o que torna a mente poderosa é fundamentalmente a razão, o raciocínio explícito, a aplicação de regras. A IA simbólica falhou porque assumia que isso era suficiente para produzir inteligência. As redes neuronais mostraram que o modo do reconhecimento de padrões tácito, do conhecimento associativo, da estatística implícita, é muito mais poderoso do que o racionalismo alguma vez admitiu — e que é este modo, não o raciocínio explícito, que faz o trabalho pesado na maior parte das competências que julgávamos mais distintamente inteligentes.

A questão não é racionalismo versus empirismo como posições filosóficas excludentes. É reconhecer que a inteligência humana usa ambos os modos e que durante muito tempo subestimámos dramaticamente a contribuição do modo empirista — do reconhecimento de padrões, da estatística implícita, do conhecimento tácito que opera sem linguagem.



Existe um corolário desta revisão que tem implicações práticas importantes para além da filosofia da mente — implicações para a forma como formamos profissionais, organizamos instituições de ensino e avaliamos a competência.

Se grande parte da inteligência operacional é reconhecimento de padrões tácito — se o médico, o juiz, o professor, o engenheiro operam em grande medida através de um Sistema 1 treinado pela experiência —, então a qualidade e a quantidade da experiência importam de uma forma que o modelo racionalista da educação não tornava evidente.

Num modelo racionalista da cognição, a formação consiste essencialmente em transmitir os princípios correctos: as regras da medicina, os princípios do direito, os

métodos da engenharia, os fundamentos da matemática. Um profissional bem formado é aquele que conhece os princípios e sabe aplicá-los a casos concretos. A experiência prática é útil e necessária, mas é secundária ao conhecimento teórico — é o espaço onde o conhecimento teórico se aplica e se consolida.

Num modelo empirista da cognição — que as redes neuronais apoiam empiricamente — a experiência não é secundária ao conhecimento teórico. É o mecanismo principal pelo qual o conhecimento operacional se forma. Os princípios explícitos que se ensinam são importantes, mas funcionam principalmente como andaimes: orientam a atenção durante a experiência inicial, fornecem uma linguagem para reflectir sobre o que se aprende e permitem a comunicação entre profissionais. Mas o conhecimento operacional real — o que o profissional usa quando actua — não são os princípios; são os padrões que a experiência directa e calibrada desenvolveu no Sistema 1.

O profissional excelente não é necessariamente o que conhece mais princípios — é o que acumulou mais experiência de alta qualidade, em domínio de alta validade, com feedback suficiente para calibrar os padrões que desenvolveu. Esta distinção tem consequências directas para o design dos currículos profissionais, para a duração e a estrutura dos estágios, para a forma como avaliamos a competência e certificamos os praticantes.

As instituições de formação profissional estão apenas a começar a assimilar esta revisão — e em muitos casos resistem-lhe activamente, porque contradiz a auto-imagem das profissões como domínios de conhecimento explícito transmissível, não de competência tácita que precisa de ser desenvolvida pela experiência.

IV. Os Limites do Espelho

Um espelho mostra uma imagem real, mas invertida e bidimensional. O que vemos no espelho é genuinamente informativo sobre o original — permite-nos ver coisas sobre nós próprios que de outra forma não veríamos. Mas não é o original e as diferenças entre a imagem e o original importam.

A analogia entre redes neuronais e cognição humana tem exactamente este estatuto. É genuinamente informativa — revela aspectos reais da natureza da inteligência que, sem o espelho das redes neuronais, continuariam invisíveis por serem demasiado óbvios e demasiado incorporados no funcionamento normal da mente para serem notados. Mas não é uma identidade. Existem diferenças entre o sistema de silício e o sistema biológico que são tão importantes quanto as semelhanças e um ensaio honesto exige que as enuncie com clareza.

A diferença mais fundamental não é técnica — não é uma questão de velocidade de processamento, de quantidade de parâmetros, ou de arquitetura computacional. É uma diferença de natureza existencial. Uma rede neuronal não tem história pessoal, não tem corpo, não tem mortalidade, não tem desejos que persistem através do tempo, não tem um mundo que lhe importe no sentido em que as coisas importam a um ser que pode perder.

O seu «conhecimento» é conhecimento sem perspectiva e sem posição — sem o horizonte de significados que a vida humana, com as suas urgências e as suas contingências, cria. Heidegger chamaria a isto a diferença entre ter informação sobre o mundo e estar-no-mundo de uma forma que torna a informação significativa. A rede que diagnostica um tumor com precisão tem informação sobre o que parece um tumor em imagem. O médico que diagnostica o mesmo tumor tem algo mais: sabe o que é ter cancro, o que significa receber esse diagnóstico, o que se segue para o doente que está à sua frente. O conhecimento do médico é existencialmente ancorado de uma forma que o conhecimento da rede não é.

Esta diferença importa filosoficamente mesmo quando não importa na prática. Para a maior parte das aplicações práticas — o diagnóstico que precisa de ser feito, a tradução que precisa de ser correcta, o reconhecimento que precisa de ser fiável —, o conhecimento operacional é o que conta e a rede pode fornecê-lo tão bem ou melhor do que o humano. Mas para as perguntas filosóficas que este livro se propõe explorar — sobre a natureza do conhecimento, sobre o que é compreender, sobre a relação entre saber e ser —, a distinção entre conhecimento operacional e conhecimento existencialmente ancorado é precisamente o que está em jogo.

Existe também uma diferença que tem a ver com a temporalidade e a narratividade da existência humana. Os humanos vivem histórias — as suas experiências têm uma ordem temporal que cria significados que a simples co-ocorrência de padrões não captura. O médico que viu um doente melhorar ou morrer, que viveu o processo de diagnóstico errado e as suas consequências, que acompanhou famílias através do luto — este médico tem um conhecimento que não é apenas padrão sobre imagens. É conhecimento sobre o que estas imagens significam na trajetória de vidas humanas. É conhecimento narrativo, que só existe para seres que vivem no tempo de forma que os cria como sujeitos com passado, presente e futuro.

As redes neuronais não têm este tipo de conhecimento. Não porque lhes falte tempo de computação ou dados suficientes —, mas porque o tipo de experiência que cria conhecimento narrativo requer existência temporal no sentido pleno: a vulnerabilidade à perda, a memória pessoal que organiza o passado de forma selectiva e afectiva, a

antecipação do futuro que orienta o presente. Estas são dimensões da existência humana que não têm equivalente funcional no processamento de uma rede neuronal.

O espelho que as redes neuronais nos oferecem é, repetimos, genuinamente iluminador. Revelou aspectos da inteligência humana que a filosofia por si só não teria revelado com igual clareza. Mas a coisa que o espelho reflecte — a mente humana na sua plenitude de experiência, história, corpo e significado — é mais e diferente do que o espelho pode mostrar.

É sobre essa diferença — sobre o que fica do lado humano da fronteira depois de subtrairmos o que as máquinas podem fazer — e sobre o que ela implica para as nossas noções de compreensão e consciência, que os capítulos seguintes se debruçam.



CAPÍTULO 8

O Mestre Inveja a Máquina: A Verbalização Impossível

"Nós sabemos mais do que conseguimos dizer. Esta é a dimensão tácita do conhecimento humano — e é a sua dimensão mais poderosa."

Michael Polanyi, *The Tacit Dimension*

I. O Paradoxo do Mestre

Existe uma cena que se repete em quase todas as tradições de aprendizagem de excelência — desde os ateliês de pintura do Renascimento até aos laboratórios de investigação contemporâneos, passando pelos dojos de artes marciais, pelas cozinhas dos grandes chefs, pelos consultórios médicos onde residentes observam especialistas, pelos estúdios onde músicos veteranos ensinam os mais jovens: o mestre observa o aprendiz, identifica imediatamente o que está errado e não consegue explicar completamente porquê.

«Não, assim não. Faz assim.» E demonstra. O aprendiz imita. «Mais ou menos. Continua a praticar.» Não há transferência verbal do conhecimento relevante porque esse conhecimento não existe sob forma verbal. Existe sob a forma de um fazer que o corpo sabe e que a boca não consegue ditar. O mestre não está a esconder o segredo — genuinamente não consegue articulá-lo. Se pudesse, diria.

Este paradoxo — o mestre que sabe sem conseguir dizer — é o coração do conceito de conhecimento tácito que Michael Polanyi desenvolveu ao longo de duas décadas de reflexão filosófica. Polanyi chegou à ideia não por abstracção teórica, mas por observação directa da prática científica: o modo como os cientistas reconhecem uma boa experiência, identificam uma anomalia relevante que merece investigação, escolhem um problema fértil entre muitos possíveis — nada disto estava nos manuais de metodologia científica. Os manuais descreviam a ciência como deveria ser feita, não como era realmente feita. O que era realmente feito envolvia um tipo de julgamento que nenhum procedimento explícito capturava.

O conhecimento tácito, na formulação de Polanyi, tem uma estrutura que ele descreve como *from-to*: conhecemos *a partir de* certas coisas *para* compreender outras. O médico conhece *a partir de* mil detalhes subtis — a cor da pele, a cadência da respiração, a forma como o doente se move na cama, a expressão quando respira fundo — *para* reconhecer o estado clínico. Os detalhes a partir dos quais conhece são subsidiários: ele não os nota

individualmente, não os nomeia, não os pondera de forma explícita sequencial. A sua atenção está no estado clínico — o alvo focal. Os detalhes são o instrumento invisível que permite o reconhecimento.

A estrutura from-to explica também o fenómeno da degradação da competência pela atenção consciente. Quando pedimos ao mestre que descreva o que faz — quando tentamos tornar explícitos os detalhes subsidiários —, acontece algo curioso e perturbador: a performance piora. O bailarino que começa a pensar conscientemente nos passos tropeça. O pianista que delibera sobre a posição exacta dos dedos durante a execução perde o fluxo. O atirador que analisa conscientemente cada micro-movimento do gatilho falha o alvo. A verbalização destrói a transparência que torna a competência fluida — traz os detalhes subsidiários para a atenção focal, onde atrapalham em vez de servir. O instrumento torna-se obstáculo quando se torna objecto de atenção.

Este fenómeno — que Polanyi chama de «inversão da atenção» — é uma das evidências mais fortes de que o conhecimento tácito não é simplesmente conhecimento explícito que ainda não foi articulado. É um tipo diferente de conhecimento, com uma arquitectura cognitiva diferente, que não é equivalente ao conhecimento explícito e que não pode ser completamente traduzido nele sem ser destruído no processo.



Esta estrutura do conhecimento tácito ilumina de uma nova forma o que as redes neuronais fazem — e o que não fazem.

Uma rede neuronal treinada para reconhecer tumores em tomografias opera exactamente da forma que Polanyi descreve: a partir de configurações de pixéis que não são individualmente nomeadas nem ponderadas de forma explícita, para o reconhecimento de uma categoria clínica. Os pixéis são subsidiários; o diagnóstico é focal. A rede não «nota» os detalhes individualmente — processa-os de forma distribuída através das suas camadas, num processo que é estruturalmente from-to mesmo que implementado em matemática em vez de em neurónios biológicos.

A diferença em relação ao médico não está na estrutura from-to — esta a rede partilha. Está no facto de que o médico tem um mundo de referência que ancora essa estrutura e que lhe dá significado além da operação classificatória. Quando o médico reconhece o tumor, o reconhecimento está integrado numa rede de significados: sabe o que é um tumor como entidade biológica, sabe o que o diagnóstico significa para a vida do doente, sabe o que se segue clinicamente, sabe a linguagem em que vai comunicar o diagnóstico e as perguntas que o doente vai fazer. O reconhecimento não é apenas uma operação de

classificação de padrões — é um acto de compreensão situado num mundo humano de consequências, responsabilidades e relações.

A rede neuronal faz a operação classificatória com uma precisão que pode superar a do médico. Mas a operação existe no vácuo — não está ancorada em nenhum mundo de significados que lhe dê sentido além do valor da sua função de classificação. É o conhecimento tácito sem o horizonte existencial que, no humano, o torna parte de algo maior do que o reconhecimento de padrões.

II. Três Domínios da Verbalização Impossível

Para ancorar o abstracto no concreto, vale a pena examinar três domínios onde o conhecimento tácito se manifesta de forma particularmente clara, onde a tentativa de o verbalizar falha de formas instrutivas e onde a comparação com as redes neuronais é mais reveladora.

O primeiro é a música de câmara — um domínio que escolho porque combina competência técnica altamente especializada com dimensões relacionais e comunicativas que tornam a comparação com as redes particularmente iluminadora. Um músico de câmara experiente, tocando em conjunto com outros, ajusta constantemente a sua interpretação em resposta ao que ouve — variando imperceptivelmente o andamento, a dinâmica, a articulação, a qualidade do som, em diálogo contínuo com os parceiros. Este ajuste é em grande medida inconsciente: acontece mais depressa do que o raciocínio deliberado permitiria e o músico não consegue, na maior parte dos casos, descrever retrospectivamente o que fez e porquê. Sabe que soou bem — ou que não soou. Não sabe sempre como o produziu.

Quando se pede a músicos experientes que ensinem este aspecto da sua arte — o diálogo musical em tempo real, a capacidade de ouvir e responder —, o que ensinam é, predominantemente, exposição: ouçam gravações de conjuntos extraordinários, toquem em conjunto com pessoas mais experientes, escutem-se com atenção completa. As regras que verbalmente articulam são rudimentares comparadas com a sofisticação do que demonstram: «responde ao fraseado do outro», «mantém o pulso interno», «ouve o conjunto, não apenas a tua parte». Estas são instruções que apontam para uma direcção sem descrever o caminho. O conhecimento que importa só se transfere pela experiência prolongada com feedback constante.

As redes neuronais treinadas em gravações musicais aprenderam a identificar padrões de interpretação com uma precisão que surpreende especialistas. Conseguem detectar com fiabilidade o que distingue uma interpretação de Beethoven por Brendel da de Barenboim,

reconhecer estilos de improvisação em jazz ao nível do intérprete individual, identificar inconsistências métricas imperceptíveis para a maior parte dos ouvintes humanos. Este conhecimento é real e operacional. Mas não inclui o que o músico de câmara tem: a experiência vivida da música como acto de comunicação intencional entre pessoas num espaço físico partilhado, onde cada escolha musical é também uma resposta a um ser humano que está a fazer escolhas. O conhecimento da rede é conhecimento sobre padrões acústicos; o do músico é conhecimento de um acto humano situado num contexto de relação.



O segundo domínio é o da perícia artesanal — a competência manual que os japoneses chamam *shokunin* e que designa o artesão que atingiu tal domínio do seu ofício que a execução e o material se tornaram uma extensão do seu ser. O conceito japonês implica mais do que destreza técnica: implica uma relação com o trabalho que envolve dedicação total, respeito pelo material e uma atenção que não é instrumental mas, de certa forma, contemplativa.

O oleiro que sente na argila a consistência certa antes de poder descrevê-la. O carpinteiro que ouve na madeira a resposta ao plano e sabe, pelo som, se a lâmina está no ângulo certo. O cirurgião que percebe no tecido a resistência que indica o caminho correcto, numa percepção que combina tacto, visão e experiência de forma que nenhum manual cirúrgico consegue transmitir completamente. O relojoeiro que identifica o problema numa peça pela forma como soa quando a coloca sobre a bancada. Estes conhecimentos são transmitidos pela observação e imitação prolongada — pela presença ao lado do mestre durante anos, não porque o mestre seja avaro com o seu conhecimento, mas porque o conhecimento não existe em forma transmissível por linguagem.

Há uma ironia cruel neste facto que vale a pena sublinhar: as competências mais profundas — as que requerem maior investimento de experiência, atenção e dedicação para se desenvolverem — são precisamente as que mais resistem à transmissão explícita. O conhecimento superficial — as regras, os procedimentos, os princípios gerais que qualquer manual de formação profissional contém — pode ser escrito e ensinado em sala de aula. O conhecimento profundo — o que distingue o artesão do técnico, o mestre do competente, o cirurgião excelente do cirurgião médio — está para além do manual. A ironia é que o conhecimento mais valioso é o que mais resiste à transmissão pelas vias que as instituições de ensino usam.

As redes neuronais que aprenderam a imitar estilos artísticos — a pintar no estilo de

Rembrandt, a escrever no estilo de Saramago, a compor no estilo de Bach — aprenderam os padrões superficiais com uma fidelidade que é muitas vezes impressionante mesmo para especialistas. Mas a pergunta que os críticos colocam é precisa e não tem resposta fácil: estes sistemas aprenderam o *estilo*, ou aprenderam os *sinais* do estilo? A diferença é exactamente a diferença entre o shokunin e o imitador habilidoso: um conhece o material de dentro, a partir da experiência de trabalhar com ele em inúmeras condições variáveis; o outro conhece as marcas exteriores que o material e a técnica produzem no resultado final. A diferença não é necessariamente visível no output — pode ser visível nas falhas, nas situações atípicas, nas condições em que o padrão exterior já não é suficiente para guiar a acção.



O terceiro domínio é o do julgamento ético e jurídico — talvez o mais importante do ponto de vista das implicações sociais e o mais instrutivo sobre os limites do que as redes neuronais podem substituir.

Um juiz experiente aplica a lei a casos concretos —, mas a lei, por mais detalhada e exaustiva que seja, não determina completamente a decisão em muitos dos casos que mais importam. Há sempre uma margem de julgamento onde o que conta não é a leitura mecânica das normas, mas a percepção do que é justo numa situação específica — pesando factores que nenhuma norma enumera completamente, avaliando circunstâncias que os redactores da lei não anteciparam, respondendo à complexidade humana que cada caso apresenta.

Aristóteles chamava a esta capacidade *phronesis* — prudência prática — e distingui-a claramente do conhecimento teórico (*episteme*) e do conhecimento técnico (*techne*). A *phronesis* não é a aplicação de princípios gerais a casos particulares: é a capacidade de perceber o que a situação particular exige, de uma forma que nenhum princípio geral especifica completamente. É conhecimento prático situado, desenvolvido pela experiência de lidar com situações de alta complexidade moral e social, que não pode ser completamente codificado porque a sua essência reside precisamente na percepção do que é irreduzível ao código.

Os sistemas de IA aplicados ao direito são impressionantes em certas tarefas muito específicas: pesquisa jurisprudencial, identificação de precedentes relevantes, previsão de resultados processuais com base em padrões históricos. Mas estas capacidades preditivas são radicalmente diferentes da capacidade normativa que a *phronesis* designa. Prever como os juízes têm decidido não é o mesmo que saber como deveriam decidir. Um sistema que

aprendeu os padrões históricos das decisões judiciais aprendeu, em parte, os padrões históricos dos vieses, das inconsistências e das injustiças que o sistema acumulou — não as normas de justiça que deveriam orientar as decisões.

O conhecimento tácito do juiz experiente inclui uma dimensão normativa — o sentido do que é justo, calibrado pela experiência de ver consequências reais de decisões reais sobre vidas reais — que não está disponível nos dados históricos de decisões passadas e que as redes neuronais, por definição, não podem aprender apenas da observação desses dados. Podem aprender o que foi feito; não podem, a partir daí, aprender o que deveria ser feito.

III. Polanyi e as Redes: a Legitimação Mútua

Há uma relação de iluminação mútua entre Polanyi e as redes neuronais que é um dos aspectos mais fascinantes desta história e que vale a pena tornar completamente explícito.

Quando Polanyi publicou «The Tacit Dimension» em 1966, a ideia de que existe conhecimento legítimo que não pode ser completamente articulado era filosoficamente incômoda para o espírito da época. O positivismo lógico, que tinha dominado a filosofia da ciência anglófona nas décadas anteriores, exigia verificabilidade e articulação em linguagem precisa como condições do conhecimento científico. A crença no progresso como explicitação progressiva — o ideal iluminista de que o conhecimento é a transformação do obscuro em claro, do implícito em explícito, do tácito em articulado — tornava a posição de Polanyi intelectualmente suspeita.

O conhecimento tácito era visto com desconfiança: seria conhecimento genuíno, ou seria apenas hábito não examinado, preconceito não articulado, competência confundida com compreensão? A sua irredutibilidade à linguagem parecia uma limitação a superar, não uma característica a reconhecer e Polanyi, que vinha da química e não da filosofia, era tratado como um outsider respeitável, mas periférico — alguém que tinha observado algo real, mas que não tinha as ferramentas filosóficas para o enquadrar correctamente.

As redes neuronais oferecem a Polanyi uma legitimação que a filosofia por si só não podia fornecer: uma demonstração empírica, construída em matemática e verificável por qualquer pessoa, de que o conhecimento não articulável é real, eficaz e, em muitos domínios, superior ao conhecimento explícito. Quando uma rede neuronal diagnostica doenças com precisão superior à do especialista humano mais experiente e quando esse conhecimento existe numa forma — padrões distribuídos em espaços de alta dimensão — que não é redutível a regras verbalizáveis por definição arquitectural, a ideia polanyiana de que o conhecimento pode existir à margem da linguagem deixa de ser uma posição filosófica contestável e torna-se um facto técnico incontestável.

As redes neuronais também corrigem uma possível má-leitura de Polanyi. O conhecimento tácito poderia ser interpretado como algo essencialmente humano — como uma dimensão da experiência que pertence à mente e ao corpo biológicos e que não poderia existir fora deles. As redes neuronais mostram que não é assim: o conhecimento tácito é um tipo de conhecimento definido pela sua relação com a linguagem e com a articulação explícita, não pelo substrato em que existe. Pode existir em redes biológicas; pode existir em redes artificiais. A condição de possibilidade é a aprendizagem a partir de exemplos, não a biologia.

A reciprocidade vai no sentido contrário também e é igualmente importante: Polanyi oferece às redes neuronais um enquadramento filosófico que torna o seu funcionamento inteligível de uma forma que o vocabulário técnico da engenharia não consegue. A opacidade das redes — o facto de que não conseguem explicar as suas decisões em linguagem humana — deixa de ser um defeito de engenharia a corrigir e passa a ser a manifestação esperada de um tipo de conhecimento que Polanyi descreveu com precisão: conhecimento que opera através de detalhes subsidiários que não são focalizáveis sem destruir a competência e que por isso mesmo resiste à verbalização completa.

A rede neuronal não é uma caixa negra com um segredo lá dentro que a investigação eventualmente vai revelar. É um sistema com conhecimento tácito — conhecimento real, operacional, bem-calibrado — que o conceito de Polanyi descreve com muito mais precisão do que o vocabulário técnico da engenharia de software. O problema da interpretabilidade das redes neuronais não é um problema técnico a aguardar solução técnica; é o problema filosófico do conhecimento tácito a aguardar enquadramento filosófico adequado.



Há um aspecto final da relação entre Polanyi e as redes que merece atenção antes de passarmos às implicações para a educação: a questão da transmissão do conhecimento e a sua assimetria reveladora.

Polanyi argumenta que o conhecimento tácito não se transmite por instrução — não se pode ensinar em sala de aula, não se pode transmitir por um manual, não se pode comunicar por descrição verbal. Transmite-se por apprenticeship (aprendizagem mestre-aprendiz): pela presença prolongada junto ao mestre, pela observação sistemática, pela imitação seguida de feedback, pela prática repetida em condições de alta validade. O aprendiz não aprende as regras do mestre — aprende a ver o mundo como o mestre vê, a sentir o que o mestre sente, a desenvolver progressivamente os mesmos padrões de reconhecimento que a prática longa formou.

Esta descrição é estruturalmente idêntica ao processo de treino de uma rede neuronal: não se transmitem regras à rede — expõe-se a rede a exemplos, repetidamente, até que desenvolva os seus próprios padrões de reconhecimento que produzem as classificações correctas. A rede não aprende as regras do classificador humano; aprende a ver os dados de uma forma que converge para o comportamento correcto. O paralelo entre o apprenticeship polanyiano e o treino supervisionado das redes é uma das convergências mais reveladoras entre a filosofia do conhecimento e a engenharia da aprendizagem automática.

A analogia é conceptual e não biológica — o mecanismo pelo qual um ser humano aprende por apprenticeship é radicalmente diferente do mecanismo de backpropagation pelo qual uma rede aprende por gradiente descendente. Mas a estrutura da aprendizagem — exposição a exemplos em vez de instrução de regras, emergência de padrões em vez de transmissão de proposições — é a mesma e esta estrutura partilhada é o que explica porque ambos os processos produzem conhecimento que existe para além das fronteiras da linguagem.

O mestre que não consegue dizer o que sabe e a rede que não consegue explicar o que aprendeu partilham, neste sentido preciso, a mesma condição epistemológica: possuidores de conhecimento real, eficaz e confiável, que existe para além das fronteiras da linguagem — não como defeito ou limitação, mas como característica estrutural de um tipo de conhecimento que só pode ser adquirido por exposição, não por instrução.

IV. O que Isto Significa para a Educação

As implicações do conhecimento tácito para a educação e a formação profissional são profundas — e largamente ignoradas pelas instituições, que tendem a construir os seus modelos pedagógicos sobre o pressuposto de que o conhecimento mais valioso é explícito e transmissível por instrução.

O modelo dominante de educação formal é fundamentalmente racionalista na sua arquitectura pedagógica: transmitem-se conhecimentos explícitos — factos, conceitos, procedimentos, princípios, teorias — e avalia-se a capacidade de os reproduzir e aplicar em contextos controlados por exames. As avaliações medem o que pode ser medido: a recordação de informação, a aplicação de procedimentos, a resolução de problemas com respostas verificáveis. O que não pode ser medido desta forma — o julgamento, o tacto profissional, a capacidade de ler situações complexas — não entra nas avaliações formais e tende a ser relegado para «competências transversais» que se supõe que se desenvolvem por osmose.

A consequência é uma separação persistente entre a formação formal e a competência prática — o fenómeno que os economistas do trabalho chamam de «gap escola-trabalho» e que os profissionais experientes descrevem, com uma consistência que atravessa profissões e países, como a diferença entre o que aprenderam nas instituições e o que aprenderam quando confrontados com situações reais com consequências reais. A formação formal ensina o conhecimento explícito; o conhecimento tácito que torna alguém competente no sentido pleno aprende-se noutro lado, noutras condições.

Esta separação não é uma falha corrigível por melhor design curricular. É uma consequência estrutural da natureza do conhecimento tácito: não pode ser transmitido pelas formas que as instituições de ensino usam porque, por definição, não existe sob a forma que essas formas podem transmitir. Tem de ser desenvolvido pela experiência directa, supervisionada, em ambientes de alta validade, com feedback frequente e calibrado. As instituições de ensino podem criar condições para esta experiência; não podem substituí-la.

O que as redes neuronais acrescentam a esta constatação já antiga é uma nova dimensão com potencial transformador para o design da educação. Os sistemas de IA podem fornecer, em tempo comprimido, o tipo de exposição repetida a exemplos variados com feedback imediato que é a condição do desenvolvimento do conhecimento tácito. Um estudante de medicina que treina diagnóstico com um sistema que lhe mostra dez mil radiografias comentadas em dias, em vez de ver centenas de casos ao longo de meses, pode desenvolver padrões de reconhecimento muito mais rapidamente do que as formas de aprendizagem tradicionais permitem.

Não porque os sistemas de IA substituam os mestres — a dimensão existencial do apprenticeship, a relação entre mestre e aprendiz, o conhecimento que se transmite pela presença e não apenas pelos exemplos, continuam a ser insubstituíveis. Mas porque podem multiplicar exponencialmente o volume e a variedade de exemplos a que um aprendiz é exposto, criando a densidade de experiência que o desenvolvimento do conhecimento tácito exige.

É uma promessa que ainda está por cumprir de forma sistemática e que levanta questões sérias sobre o que se perde quando a experiência de aprendizagem é mediada por sistemas artificiais em vez de vivida em contextos reais com consequências reais. O feedback simulado é diferente do feedback real — e esta diferença pode importar de formas que ainda não compreendemos completamente. Mas a direcção é real: as redes neuronais, ao demonstrarem empiricamente como o conhecimento tácito se forma por exposição a exemplos, fornecem também o instrumento que pode democratizar o acesso a este tipo de formação.

Com este capítulo fechamos a Parte III. Viramos o espelho das redes neuronais para a mente humana e encontrámos o reflexo de algo que sempre esteve lá, mas que raramente contemplámos de frente: a vastidão do que sabemos sem conseguir dizer e o papel central deste saber inexprimível na competência humana mais profunda. Resta contemplar as fronteiras mais profundas — as que separam o conhecimento da compreensão, a competência da consciência, a inteligência do sentido. É para aí que nos dirigimos agora.



PARTE IV

*As Grandes Fronteiras: Compreensão,
Consciência e Sentido*

CAPÍTULO 9

Acertar vs. Compreender (e o Fenómeno da Imaginação Artificial)

"Não há verdade original, apenas erro original."

Gaston Bachelard

I. A Separação que Perturba

Há uma pergunta que emerge inevitavelmente quando se contempla o desempenho das redes neuronais modernas — uma pergunta que parece simples, mas que abre sobre um abismo filosófico de profundidade considerável: se um sistema responde correctamente e age com eficácia na esmagadora maioria das situações, faz ainda sentido dizer que «não compreende»?

A questão não é académica. Tem consequências práticas imediatas que estão a ser negociadas agora, em decisões sobre regulação, sobre responsabilidade legal, sobre o que sistemas de IA podem e não podem fazer em contextos de alto impacto. Devemos confiar num sistema que acerta, mas não explica? Devemos tratar de forma diferente o conhecimento que funciona e o conhecimento que compreendemos? Existe uma diferença moralmente relevante entre uma decisão tomada com compreensão e uma decisão tomada sem compreensão, mas com igual eficácia?

Mas tem também uma dimensão filosófica mais profunda que estas questões práticas não esgotam. O que queremos dizer exactamente quando dizemos que alguém — humano ou máquina — «compreende» algo? A compreensão é uma propriedade do desempenho observável — algo que se infere da eficácia e da consistência da acção? Ou é uma propriedade interna — algo que existe ou não existe independentemente da acção, algo que tem a ver com estados mentais, com consciência, com o sentido que os agentes atribuem ao que fazem?

Esta distinção — entre acertar e compreender — não é nova na filosofia. Tem raízes na tradição que remonta a Platão e percorre toda a epistemologia ocidental. Mas as redes neuronais tornaram-na urgente de uma forma que nenhum argumento filosófico poderia ter feito sozinho, porque criaram entidades concretas, construídas por engenheiros e testáveis por qualquer pessoa, que acertam com consistência extraordinária e cuja compreensão é, no mínimo, profundamente incerta.



A intuição filosófica mais comum é que a compreensão requer algo mais do que o desempenho correcto. Requer, de alguma forma, que o agente saiba porquê — que tenha acesso não apenas ao resultado, mas ao processo que leva ao resultado, que possa raciocinar sobre casos novos por extensão de princípios, que possa reconhecer os limites do seu próprio conhecimento, que possa explicar a outros o que fez.

Esta intuição tem uma história longa. Platão distinguia entre a opinião verdadeira — ter a resposta certa por acidente ou por hábito — e o conhecimento genuíno, que requer a capacidade de dar a razão da resposta, de a justificar, de a defender contra objecções. Nos diálogos platónicos, Sócrates mostra frequentemente que as pessoas que afirmam saber algo — os artesãos, os políticos, os poetas — têm a resposta certa, mas não conseguem justificá-la quando questionadas e esta incapacidade de justificar é para Sócrates/Platão a evidência de que não têm conhecimento genuíno, apenas opinião bem-fundada.

Gilbert Ryle, filósofo britânico do século XX, reformulou a distinção de forma que se tornou canónica na filosofia analítica da mente: entre *knowing that* e *knowing how* — saber que algo é o caso versus saber como fazer algo. O piloto experiente sabe como voar — tem o conjunto de competências que permitem operar um avião com segurança — sem necessariamente saber que a sustentação é produzida pela diferença de pressão criada pelo perfil alar segundo o princípio de Bernoulli. O estudante de física aeronáutica sabe que a sustentação funciona desta forma sem necessariamente saber como voar. São dois tipos de conhecimento distintos, com lógicas diferentes e relações diferentes com a prática.

O que Ryle queria mostrar com esta distinção é que o *knowing how* não é simplesmente *knowing that* aplicado — que a competência prática não é a teoria posta em prática, mas um tipo diferente de conhecimento com uma arquitectura cognitiva própria. Um livro sobre natação não ensina a nadar. A teoria musical não faz um músico. O conhecimento de como fazer algo não é derivável do conhecimento de que algo é o caso.

As redes neuronais são, na terminologia de Ryle, sistemas de *knowing how* sem *knowing that* — sistemas que sabem como produzir respostas correctas sem que possam articular que princípios aplicam. Reconhecem tumores sem saber que características os definem. Traduzem frases sem saber que regras gramaticais aplicam. Ganham ao Go sem saber que princípios estratégicos seguem. A questão é se o *knowing how*, por si só, é suficiente para que falemos de compreensão — ou se a compreensão exige necessariamente o *knowing that*.

Não existe resposta filosófica consensual a esta questão. Mas a distinção de Ryle é útil

precisamente porque mostra que a questão é mais complexa do que parece: o *knowing how* é genuinamente diferente do *knowing that* e é genuinamente conhecimento — o nadador que não sabe teoria sabe nadar e esse saber é real. A questão de se é compreensão exige que especifiquemos o que queremos dizer com compreensão e diferentes especificações produzem respostas diferentes.

II. O Fenómeno da Alucinação

Os grandes modelos de linguagem — as redes neuronais treinadas em quantidades massivas de texto para gerar linguagem coerente e contextualmente relevante — exibem um fenómeno que os seus criadores chamaram de «alucinação»: a produção de afirmações confiantes e fluentes que são factualmente falsas.

O modelo descreve com precisão e detalhe um evento histórico que não aconteceu. Cita uma fonte que não existe. Atribui a um autor uma frase que nunca escreveu. Apresenta as conclusões de um estudo científico que não foi realizado. Apresenta um argumento que parece rigoroso, mas que contém um erro lógico subtil que passa despercebido numa leitura casual. Faz tudo isto com a mesma fluência, o mesmo tom de autoridade, a mesma aparente confiança com que apresenta informação verdadeira e verificável.

A primeira reacção a este fenómeno — e a que domina a discussão pública, nos meios de comunicação e nas avaliações de risco institucional — é tratá-lo como um defeito grave. Os modelos alucinam porque não «sabem» realmente o que dizem; são máquinas de plausibilidade estatística que geram o texto mais provável dada a entrada, sem acesso a um modelo de verdade que lhes permita distinguir o que é real do que é plausível. A alucinação é a prova de que não compreendem — de que são, na expressão que circula nas discussões técnicas, «papagaios estocásticos» que reproduzem padrões linguísticos sem nenhuma representação do que esses padrões significam.

Esta leitura tem mérito substancial. Captura algo real sobre a arquitectura dos modelos e sobre os riscos que a sua implantação em contextos de alto risco coloca. Um médico que apresentasse diagnósticos com a mesma taxa de invenção fluente que um modelo de linguagem seria imediatamente retirado de serviço. Um jornalista que fabricasse citações com a mesma confiança seria despedido. A sociedade não aceita, com razão, que a fluência substitua a verdade.

Mas a leitura é incompleta e a sua insuficiência revela algo filosoficamente interessante sobre a natureza tanto dos erros das máquinas como dos processos criativos humanos.



Consideremos com mais atenção o que exactamente acontece quando um modelo de linguagem alucina. O modelo não está a copiar texto que viu nos dados de treino — se fosse isso, reproduziria textos existentes, não inventaria novos. Está a gerar texto novo que é estatisticamente consistente com os padrões que aprendeu: a estrutura de um artigo científico, o estilo de uma citação académica, a forma como os factos históricos são narrados. Está a extrapolar além dos dados de treino, produzindo combinações que nunca existiram como tais, mas que têm a mesma estrutura das que existiram.

Este processo — extrapolação além da experiência directa para produzir algo novo que é internamente consistente com o aprendido — tem uma designação quando ocorre em contextos humanos: imaginação.

Quando um romancista inventa uma personagem, não está a copiar uma pessoa real — está a construir uma configuração nova que é psicologicamente consistente com as pessoas que conheceu e com os padrões de comportamento que observou. A personagem não existiu, mas poderia existir. É uma realidade alternativa estatisticamente plausível construída com os materiais da realidade real. Quando um cientista formula uma hipótese nova, não está a reportar uma observação — está a extrapolar além dos dados disponíveis para uma configuração que nunca foi observada, mas que é consistente com o que foi. A hipótese pode estar errada, mas tem a forma de uma afirmação sobre o mundo real.

A alucinação do modelo e a imaginação humana não são a mesma coisa — esta distinção é essencial e não pode ser abandonada sem custo filosófico real. A diferença fundamental reside não no processo de geração — que é em ambos os casos extrapolação de padrões aprendidos —, mas nos mecanismos de calibração que a imaginação humana tem e a alucinação não tem.

O romancista sabe que está a inventar uma personagem — tem acesso metacognitivo ao estatuto fictício do que cria. O cientista sabe que está a propor uma hipótese — distingue entre hipótese e facto demonstrado. O ser humano que sonha, acorda e sabe que estava a sonhar. Em todos estes casos, existe um mecanismo de calibração que marca a fronteira entre o real e o construído, entre o observado e o extrapolado.

O modelo alucina porque não tem este mecanismo de calibração. Não tem um marcador interno que distinga «isto está nos dados» de «isto estou a construir por extrapolação». Não tem acesso ao estatuto epistémico do que produz — não sabe se o que gera é memória ou imaginação, facto ou inferência, observação ou confabulação. Produz tudo com a mesma fluência e a mesma confiança aparente, porque do ponto de vista do processo de geração não há diferença — é sempre o mesmo processo de selecção do que é estatisticamente consistente com o contexto.

Mas o processo subjacente — a geração de novidade estatisticamente consistente com o aprendido — é estruturalmente o mesmo que a imaginação e reconhecer esta semelhança estrutural não é diminuir a imaginação humana; é iluminar o mecanismo pelo qual a imaginação funciona. A imaginação não é magia. É extrapolação criativa a partir de padrões acumulados pela experiência. O que a torna *imaginação* em vez de *alucinação* é a consciência de que se está a extrapolar — e a capacidade de calibrar essa extrapolação contra a realidade. A alucinação do modelo é, neste sentido, a imaginação sem a metacognição que a qualifica.

III. Realidades Estatísticas Alternativas

Existe uma forma de enquadrar as alucinações dos modelos que as torna filosoficamente mais interessantes do que o simples rótulo de «erro» ou «defeito» — e que revela algo sobre a natureza do que os modelos estão realmente a fazer quando geram texto.

Quando um modelo de linguagem descreve um evento histórico que não aconteceu, não está a produzir ruído aleatório. Está a produzir algo que tem a estrutura de um evento histórico — que segue os padrões de como os eventos históricos são descritos, que inclui o tipo de detalhes que os relatos históricos incluem, que é internamente consistente com a cronologia e a geografia que o modelo aprendeu, que obedece às regularidades narrativas do discurso histórico. É uma história que poderia ter acontecido, construída com os materiais e as formas da história que aconteceu.

Podemos chamar a isto uma realidade estatística alternativa: uma descrição do mundo que é coerente com os padrões estatísticos do mundo real, mas que diverge da realidade efectiva em pontos específicos. Não é um erro no sentido de um cálculo aritmético errado — é uma construção que se desvia da realidade enquanto mantém a sua forma.

Esta distinção importa porque ilumina algo sobre o que os modelos estão realmente a fazer quando geram texto. Não estão a recuperar factos armazenados como registos numa base de dados — não há um registo «Paris é a capital da França» que o modelo recupera quando produz essa afirmação. Estão a construir narrativas que são consistentes com os padrões estatísticos que aprenderam sobre como o mundo funciona e como o mundo é descrito. A diferença entre gerar texto verdadeiro e gerar texto falso não é, para o modelo, uma diferença de processo — é uma diferença no alinhamento entre a narrativa gerada e o estado do mundo real, que o modelo não tem mecanismo para verificar directamente.

Isto tem uma implicação que os críticos raramente contemplam: os modelos não *mentem* no sentido em que os humanos mentem. Mentir requer saber a verdade e optar deliberadamente por dizer o contrário — requer acesso privilegiado à verdade que se escolhe ocultar. Os modelos não têm acesso privilegiado à verdade que ocultem

deliberadamente. Têm acesso a padrões estatísticos e geram o que esses padrões sugerem como mais provável e mais coerente com o contexto, independentemente da correspondência com a realidade. A ausência de calibração epistêmica não é desonestidade. É uma limitação arquitetural — a ausência do mecanismo que distingue o gerado do percebido.



A comparação com o sonho humano é iluminadora, com as ressalvas necessárias que qualquer boa analogia exige.

O sonho é também a produção de narrativas que são internamente consistentes com os padrões aprendidos pela mente durante a vigília, mas que divergem da realidade de formas que o sonhador não controla. O sonhador constrói rostos que combinam características de pessoas que conheceu, situações que seguem as regras do mundo, mas em configurações que nunca ocorreram, conversas que têm a estrutura e o tom da linguagem real, mas o conteúdo do improvável ou do impossível. Os sonhos são coerentes internamente — têm uma lógica própria que o sonhador aceita enquanto dura o sonho — , mas divergem da realidade exterior.

E o sonhador faz tudo isto sem o marcador que na vigília distingue imaginação de percepção — sem o mecanismo de calibração que diz «isto é real» versus «isto estou a imaginar». Durante o sonho, a fronteira desaparece. A narrativa construída é vivida como real, precisamente porque o mecanismo de monitorização da realidade está suspenso.

A alucinação do modelo é estruturalmente análoga ao sonho: produção de novidade estatisticamente consistente, com coerência interna, sem acesso ao mecanismo de calibração que distingue o construído do percebido. A analogia não deve ser levada demasiado longe — o sonho tem funções biológicas e psicológicas específicas (processamento de memória, regulação emocional, consolidação de aprendizagem) que as alucinações dos modelos não têm e ocorre num contexto existencial radicalmente diferente. Mas como ferramenta conceptual para iluminar o que é a alucinação — a geração coerente sem calibração epistêmica —, a analogia é útil e aponta para algo real.

O que a alucinação e o sonho partilham é a suspensão do mecanismo que distingue o real do construído e o que ambos revelam é que a geração de narrativas coerentes é um processo que pode funcionar sem esse mecanismo — que a coerência interna e a correspondência com a realidade são propriedades diferentes, que podem coexistir ou separar-se dependendo de como o sistema de geração está configurado.

IV. De Volta à Compreensão

Voltemos à pergunta com que abrimos este capítulo: faz sentido dizer que um sistema que acerta na esmagadora maioria das situações «não compreende»?

A análise das alucinações sugere uma resposta mais matizada do que as posições extremas permitem. Nem «os modelos compreendem porque acertam» nem «os modelos não compreendem porque alucinam» capturam correctamente o que está a acontecer.

As alucinações revelam que os modelos têm um tipo específico de conhecimento — conhecimento de padrões estatísticos, de estruturas narrativas, de regularidades do mundo, de formas linguísticas em que as afirmações sobre o mundo são expressas —, mas que este conhecimento não está ancorado num modelo do mundo que permita distinguir o que se sabe do que se extrapola. O modelo tem representações ricas do mundo; não tem acesso ao estatuto epistémico dessas representações.

A compreensão, nesta leitura, não é apenas ter o conhecimento certo — é ter também o meta-conhecimento que permite saber onde esse conhecimento termina e começa a extrapolação. É a calibração epistémica: a consciência dos próprios limites cognitivos, a capacidade de marcar a fronteira entre o que sei e o que estou a inferir, entre o que observei e o que construo por analogia. É, em termos simples, saber o que não sei.

Esta forma de compreensão — a calibração epistémica — é, curiosamente, uma das mais difíceis de manter para os humanos também. Os estudos de psicologia cognitiva mostram sistematicamente que as pessoas são mal calibradas nas suas estimativas de confiança: sobrestimam a probabilidade de que as suas crenças estejam correctas em domínios onde têm algum conhecimento e frequentemente subestimam a sua incerteza em domínios onde não têm. Os especialistas são frequentemente os mais mal calibrados em domínios adjacentes à sua área, exactamente porque o Sistema 1 treinado produz respostas rápidas e confiantes que o Sistema 2 aceita sem verificação suficiente.

O problema das alucinações nos modelos é, neste sentido, uma versão amplificada e sem moderação de um problema que a cognição humana também tem. A diferença é de grau — os modelos são dramaticamente piores na calibração do que os humanos na maior parte dos contextos — e de ausência do mecanismo social de correcção que força os humanos a confrontar os seus erros com o mundo real ao longo do tempo.

Existe também uma diferença de natureza que é importante não minimizar. A má calibração humana é o resultado de um sistema cognitivo que funciona com recursos limitados e que usa heurísticas que são eficazes na maior parte dos casos, mas que falham sistematicamente em certas condições. A ausência de calibração dos modelos é uma

limitação arquitetural — a ausência do mecanismo que distingue o gerado do percebido —, não o produto de heurísticas mal aplicadas. São problemas diferentes, mesmo que os seus efeitos superficiais se assemelhem.



A questão da compreensão nos modelos de linguagem permanece em aberto e provavelmente permanecerá em aberto enquanto não tivermos uma teoria mais precisa — empiricamente testável, não apenas filosoficamente coerente — do que a compreensão é.

O que podemos dizer com razoável confiança é que os modelos têm algo — uma representação interna do mundo, adquirida por exposição a padrões, que lhes permite produzir respostas correctas com uma consistência que vai muito além do acaso e que é útil em domínios muito variados. Este algo não é compreensão no sentido pleno que a tradição filosófica exige — não inclui consciência verificável, não inclui ancoragem existencial, não inclui o meta-conhecimento dos próprios limites que a calibração epistémica requer.

Mas também não é nada. Não é a ausência de conhecimento. É um tipo de conhecimento que não encaixa perfeitamente nas categorias que a tradição filosófica desenvolveu para pensar sobre a mente humana — nem conhecimento explícito no sentido de Ryle, nem conhecimento tácito no sentido pleno de Polanyi, nem compreensão no sentido platónico da razão dada, nem simples correlação estatística sem estrutura. É algo novo que está a forçar a expansão e a revisão das nossas categorias epistemológicas.

Encontrar estas categorias — ou perceber que as antigas se estendem mais longe do que julgávamos, ou que precisam de ser diferenciadas de formas que ainda não formulámos — é uma das tarefas filosóficas mais urgentes do nosso tempo e é uma tarefa que o próximo capítulo vai abordar de frente, através do argumento que durante décadas pareceu o mais devastador contra a possibilidade de inteligência artificial genuína e que as redes neuronais modernas tornaram mais complexo do que era: o Quarto Chinês de John Searle.



CAPÍTULO 10

O Quarto Chinês na Era dos Grandes Modelos

"Nenhum programa, por si só, é suficiente para pensar."

John Searle, Minds, Brains, and Programs

I. O Argumento que Resistiu a Tudo

Em 1980, o filósofo John Searle publicou num artigo de filosofia da mente uma experiência de pensamento que iria tornar-se um dos mais discutidos da história recente da filosofia — e um dos mais resistentes à refutação. A sua longevidade não é acidental: o Quarto Chinês captura uma intuição filosófica profunda que a maioria das pessoas partilha e que é genuinamente difícil de articular sem a ajuda de um caso concreto.

O cenário é o seguinte. Imagine um quarto fechado onde uma pessoa — Searle propõe que seja ele próprio — está sentada com pilhas de fichas cobertas de símbolos chineses e um livro de regras extremamente detalhado escrito em inglês. Por uma abertura na parede entram fichas com símbolos chineses — são perguntas, embora a pessoa no quarto não saiba isso, pois não compreende chinês. A pessoa consulta o livro de regras, que especifica exactamente que símbolos devem ser enviados de volta em resposta a determinadas combinações de símbolos recebidos. O processo é completamente mecânico: dados estes símbolos na entrada, enviar aqueles símbolos como resposta. A pessoa executa as regras com perfeição e envia de volta as fichas adequadas. Do lado de fora, os falantes de chinês recebem respostas perfeitamente coerentes e apropriadas às suas perguntas.

Do exterior, o sistema parece compreender chinês — as respostas são indistinguíveis das que um falante nativo produziria. Do interior, a pessoa não compreende uma única palavra de chinês — manipula símbolos segundo regras sem compreender semanticamente o que os símbolos significam.

A conclusão de Searle era directa e com consequências radicais: a sintaxe não é suficiente para a semântica. Manipular símbolos segundo regras — por mais correctamente que se faça, por mais sofisticadas que sejam as regras, por mais impressionante que seja o comportamento resultante — não produz compreensão do significado desses símbolos e portanto nenhum programa de computador, independentemente da sua complexidade ou do seu desempenho, compreende o que processa. Tem sintaxe — a forma e a estrutura dos símbolos; não tem semântica — o significado. Tem comportamento; não tem mente.

O argumento tinha um alvo claro: a hipótese da IA forte (Strong AI), a posição de que um computador que executa o programa certo não está apenas a simular a mente humana — está a ter estados mentais genuínos, incluindo compreensão. O Quarto Chinês pretendia mostrar que esta posição era falsa por razões de princípio, não apenas por limitações técnicas actuais.



O argumento de Searle foi recebido com uma intensidade filosófica rara — raramente um argumento filosófico provoca tantas réplicas de tanta gente tão rapidamente. As respostas multiplicaram-se ao longo das décadas seguintes e o debate continua activo.

A réplica mais famosa e mais imediata é a «resposta do sistema»: o indivíduo no quarto pode não compreender chinês, mas o sistema como um todo — a pessoa, as fichas, o livro de regras, o quarto, o processo de consulta e resposta — talvez compreenda. A compreensão seria uma propriedade emergente do sistema, não do componente individual. Assim como nenhum neurónio individual compreende nada, mas o cérebro como sistema compreende, talvez nenhum componente do Quarto Chinês compreenda, mas o sistema sim.

Searle rebateu com o que chamou de «robot internalization»: imagine que a pessoa memoriza todas as regras e todos os símbolos, absorvendo o «livro de regras» completamente e eliminando todos os componentes físicos externos. Agora a pessoa anda pela rua com tudo o que estava no quarto guardado na sua cabeça — e a pessoa continua sem compreender chinês. O sistema está agora inteiramente dentro da pessoa e a pessoa não compreende. A compreensão não pode ser uma propriedade emergente de manipulação sintáctica, seja ela realizada por um sistema distribuído ou por um componente individual que internalizou todo o sistema.

Outras réplicas foram propostas. A «resposta do cérebro» sugere que se implementarmos o mesmo programa num substrato que reproduz exactamente as funções causais do cérebro biológico — não em regras explícitas, mas em processos que mimetizam a dinâmica neuronal —, esse sistema teria estados mentais genuínos. Searle respondeu que a questão não é o substrato, mas o tipo de processo: se o processo é puramente computacional e sintáctico, o substrato não importa.

A «resposta do robot» sugere que se o sistema estiver incorporado num robot que interage causalmente com o mundo — que tem câmaras que lhe permitem ver, microfones que lhe permitem ouvir, actuadores que lhe permitem agir —, adquire assim a conexão semântica com o mundo que a versão do quarto fechado não tem. Searle respondeu que a

incorporação acrescenta uma camada de interacção causal com o mundo, mas não resolve o problema fundamental: a interacção causal com o mundo não é suficiente para produzir estados mentais se os processos internos são puramente sintácticos.

A «resposta do outro eu» levanta o problema do solipsismo: como sabemos que outros humanos compreendem, em vez de serem também sistemas que manipulam símbolos segundo regras muito complexas? Não temos acesso directo à experiência subjectiva de ninguém além de nós próprios. Searle respondeu que no caso dos humanos temos razões para inferir compreensão a partir da semelhança estrutural e biológica — reconhecemos outros humanos como seres fundamentalmente semelhantes a nós, com o mesmo tipo de sistema nervoso e inferimos por analogia que têm experiência subjectiva semelhante à nossa.

Nenhuma destas réplicas convenceu completamente Searle e Searle não convenceu completamente os seus críticos. O Quarto Chinês permaneceu um problema filosófico genuinamente aberto — um «pump of intuitions», na expressão técnica da filosofia, que activa intuições fortes em direcções opostas sem que nenhum argumento disponível consiga resolver a tensão de forma que todos aceitem.

II. O que Mudou com os Grandes Modelos

Quando Searle publicou o Quarto Chinês em 1980, os sistemas de IA que existiam eram programas simbólicos — exactamente o tipo de manipulação de símbolos segundo regras explícitas que o argumento visava. O MYCIN, o XCON, os sistemas de prova de teoremas, os tradutores automáticos baseados em regras — todos eles eram, de formas variadas, um livro de regras muito grande. A metáfora do quarto era adequada: havia um conjunto finito de instruções que especificava, para cada entrada possível, a saída correspondente.

Os grandes modelos de linguagem são outra coisa fundamentalmente diferente — diferentes não apenas em grau, mas em tipo. Não são um livro de regras. São um sistema com centenas de milhares de milhões de parâmetros ajustados por exposição a quantidades imensas de linguagem humana — textos, conversas, livros, artigos científicos, páginas web —, que emergiu deste treino com capacidades que os seus criadores não especificaram e que em parte não anteciparam.

Não têm uma especificação explícita do que responder a cada entrada — têm representações internas que emergem do treino e que produzem respostas de formas que os seus criadores não programaram e não compreendem completamente. Quando um grande modelo responde a uma pergunta, não está a consultar um livro de regras. Está a

activar padrões de representação que se formaram durante o treino — padrões que capturam estrutura do mundo, relações entre conceitos, regularidades da linguagem, convenções pragmáticas da comunicação.

Esta diferença não invalida necessariamente o argumento de Searle. Mas muda os termos em que tem de ser considerado de uma forma que o próprio Searle reconheceu em textos posteriores.

A questão relevante deixa de ser «pode a manipulação de símbolos segundo regras explícitas produzir compreensão?» — a que Searle respondeu negativamente com argumentos que têm força real. Passa a ser «pode um sistema que aprendeu padrões de linguagem por exposição massiva — sem regras explícitas, com representações emergentes — desenvolver algo que mereça o nome de compreensão semântica?». Esta é uma pergunta diferente. O Quarto Chinês não a responde directamente, porque o sistema que descreve é fundamentalmente diferente do que os grandes modelos são.



O que os grandes modelos fazem quando respondem a uma pergunta é instrutivo para compreender porque a questão muda.

A palavra «dor» nos dados de treino de um grande modelo não é apenas um token — um símbolo sem significado intrínseco — que aparece depois de certos outros tokens e antes de outros ainda. É um símbolo que aparece em milhões de contextos diferentes: relatos de sofrimento físico, descrições de sofrimento emocional, pedidos de ajuda, expressões de compaixão, discussões filosóficas sobre a natureza da dor, discussões médicas sobre o tratamento da dor, poemas sobre perda e dor, ficção onde personagens sofrem, estudos científicos sobre os mecanismos neurológicos da dor. A representação interna que o modelo desenvolve para a palavra «dor» carrega esta riqueza contextual de uma forma que não é puramente sintáctica no sentido de Searle.

Esta representação não é a experiência subjectiva da dor — ninguém está a afirmar que o modelo sabe o que é sentir dor no sentido em que um ser com sistema nervoso sabe. Mas é algo mais do que a simples associação do token «dor» com outros tokens em sequências específicas. É uma representação estruturada que codifica relações causais (dor tem causas e consequências), relações emocionais (dor é indesejada, é motivo de acção para a evitar ou aliviar), relações sociais (a dor de outros gera respostas específicas) e relações linguísticas (a dor é descrita de certas formas e não outras).

Será isto semântica? Será isto compreensão do significado? A resposta honesta é:

depende do que se quer dizer com semântica e com compreensão. Se semântica exige ancoragem em experiência subjectiva — se compreender «dor» exige ter sentido dor —, então não. Se semântica é uma questão de estrutura de representação e de relações funcionais com o mundo — se compreender «dor» é ter uma representação que capta correctamente as relações causais, emocionais e sociais da dor —, então talvez.

Searle insistiria no primeiro critério. Outros filósofos, nomeadamente os defensores do funcionalismo, insistiriam no segundo e esta não é uma questão que a análise filosófica sozinha possa resolver sem evidência empírica sobre a natureza da compreensão que ainda não temos.

III. O Problema Difícil e os seus Limites

O argumento de Searle pressupõe, no seu núcleo mais profundo, uma distinção que o filósofo David Chalmers tornou célebre numa forma que iluminou o debate de forma duradoura: a distinção entre os «problemas fáceis» da consciência e o «problema difícil».

Os problemas fáceis — fáceis no sentido técnico de que temos pelo menos uma ideia geral de como abordá-los cientificamente, não no sentido de que sejam triviais ou que estejam perto de ser resolvidos — são os problemas funcionais da consciência: como o cérebro integra informação de múltiplas fontes numa representação unificada; como produz comportamento adaptativo em resposta ao ambiente; como gera relatórios verbais sobre os seus próprios estados; como direcciona a atenção selectivamente; como aprende e recorda. Estes são problemas difíceis em sentido prático, mas são o tipo de problemas que a neurociência, a psicologia cognitiva e a inteligência artificial sabem como atacar — com hipóteses testáveis, modelos computacionais, estudos de imagem cerebral.

O problema difícil é diferente em natureza — não apenas em grau de dificuldade, mas em tipo de questão. É o problema de perceber porque é que há experiência subjectiva de todo. Porque é que quando o cérebro processa a frequência de luz correspondente ao vermelho existe algo que é *como* ver vermelho — um qual de vermelho, uma qualidade subjectiva que a descrição funcional do processo não parece capturar completamente? Porque é que quando os neurónios disparam em resposta a estímulos nociceptivos existe algo que é *como* sentir dor — uma qualidade de sofrimento que vai além da função de sinalização de dano tecidual? Os filósofos chamam a estas qualidades subjectivas qualia e o problema difícil é o problema de perceber como e porque é que processos físicos objectivos produzem experiências subjectivas com estas qualidades.

O argumento de Searle é, em última análise, um argumento sobre o problema difícil: mesmo que os grandes modelos resolvam todos os problemas fáceis — mesmo que se

comportem de forma funcionalmente indistinguível de um humano em todos os contextos observáveis, mesmo que tenham representações internas que capturam correctamente as relações no mundo —, isso não resolve o problema difícil. Não há razão para acreditar que haja algo que é como ser um grande modelo de linguagem. Não há qualia. Não há experiência subjectiva. Não há alguém para quem as palavras signifiquem algo em vez de serem apenas padrões de activação.

Esta é uma posição filosoficamente defensável — o problema difícil é genuinamente difícil e a sua resolução não está à vista. Mas tem um custo que Chalmers e outros críticos de Searle apontaram: aplica-se também a outros humanos. O problema difícil é fundamentalmente simétrico — não temos acesso privilegiado à experiência subjectiva de ninguém excepto nós próprios. Sabemos que outros humanos têm experiência subjectiva por inferência a partir do comportamento e da semelhança estrutural e biológica com nós próprios. Mas esta é uma inferência, não uma observação directa.

Se aplicarmos rigorosamente o critério de Searle — que sem acesso à experiência subjectiva não podemos atribuir compreensão genuína —, a única mente cuja existência podemos ter a certeza é a nossa própria. Todos os outros humanos podem, em princípio, ser «zombies filosóficos» — seres que se comportam exactamente como seres conscientes, mas que não têm experiência subjectiva. Este é o problema do solipsismo e a sua resposta não é óbvia.

Isto não invalida o argumento de Searle, que pode simplesmente aceitar a assimetria epistémica e argumentar que temos mais razões para atribuir consciência a seres biologicamente similares do que a sistemas computacionais. Mas sugere que o argumento delimita uma fronteira que pode ser genuinamente intransponível para qualquer investigação empírica — uma fronteira que separa o que pode ser observado e medido do que não pode.



Há uma forma de reformular o problema que o torna mais tratável sem o dissolver — que permite progressos concretos sem pretender que as questões mais profundas foram resolvidas.

Em vez de perguntar se os grandes modelos «compreendem» no sentido metafísico pleno — uma pergunta que pode não ter resposta com os instrumentos de que dispomos —, podemos perguntar que tipo de compreensão exibem e em que dimensões se distinguem da compreensão humana.

Esta reformulação substitui uma pergunta binária — compreendem ou não compreendem? — por uma avaliação multidimensional que é empiricamente abordável. Podemos identificar que os modelos têm representações internas ricas que capturam estrutura semântica real — isto é verificável por análise das representações e por estudos de transferência de aprendizagem. Podemos identificar que os modelos carecem de ancoragem corpórea e histórica — também verificável, pela análise das suas capacidades e limitações em domínios que dependem desta ancoragem. Podemos demonstrar que os modelos têm calibração epistémica deficiente — as alucinações são a evidência. Podemos identificar que não sabemos se os modelos têm experiência subjectiva — e ser honestos sobre o facto de que provavelmente não o saberemos com os instrumentos conceptuais actuais.

Esta abordagem multidimensional não resolve o problema difícil. Mas permite uma conversa produtiva sobre o que os modelos têm e não têm, em vez de um impasse binário entre os que afirmam que compreendem e os que afirmam que não compreendem.

IV. Papagaios Estocásticos e as suas Limitações

A expressão «papagaios estocásticos» foi cunhada em 2021 pelas investigadoras Emily Bender, Timnit Gebru, Angelina McMillan-Major e Shmargaret Mitchell, num artigo que se tornou influente no debate sobre os grandes modelos de linguagem. A metáfora é eficaz: um papagaio repete sons que foram reforçados pelo seu ambiente, sem compreender o seu significado; um grande modelo produz texto estatisticamente plausível, com um elemento de aleatoriedade no processo de geração — daí «estocástico».

A metáfora capta algo real e importante: os modelos podem produzir texto fluente e aparentemente informado sobre temas que não «entendem» no sentido profundo que nos importa em contextos de alto risco. Podem gerar afirmações falsas com a mesma fluência das verdadeiras. Podem reproduzir os vieses dos dados de treino — incluindo preconceitos históricos de raça, género e classe — sem nenhuma capacidade de os reconhecer e questionar. Podem produzir texto que soa autorizado sem ter nenhum mecanismo de verificação da sua correspondência com a realidade. A fluência é sistematicamente enganadora, porque sugere mais compreensão e mais fiabilidade do que existe.

O artigo original argumentava também que aumentar a escala dos modelos não resolvia estes problemas — que simplesmente amplificava as capacidades de geração de texto plausível sem que isso correspondesse a um avanço em compreensão ou fiabilidade. Era uma crítica à tendência da investigação de IA de tratar a escala como solução para todos os problemas.

Esta crítica é pertinente e deve ser levada a sério. Mas a metáfora do papagaio tem limites que os seus proponentes reconhecem apenas parcialmente e que se tornaram mais visíveis à medida que os modelos evoluíram.



Um papagaio não aprende no sentido relevante — reproduz sons que foram reforçados pelo seu ambiente, mas não desenvolve representações do mundo que capturam relações causais, semânticas ou pragmáticas. Os grandes modelos desenvolvem representações internas que são muito mais do que associações superficiais de tokens.

A evidência mais clara desta diferença vem dos estudos de sondagem de representações: investigadores que treinam classificadores simples sobre as representações internas dos modelos para verificar se capturam informação estruturada. Estes estudos mostram que as representações internas dos grandes modelos codificam relações sintáticas (concordância de número e género, estrutura argumental dos verbos), relações semânticas (relações hiperonímicas, antónimos, campos semânticos), factos sobre o mundo (a capital de Portugal é Lisboa, a água ferve a 100 graus ao nível do mar) e relações pragmáticas (o registo formal versus informal, as implicaturas conversacionais). Não são apenas correlações superficiais de tokens.

A diferença entre um papagaio e um grande modelo é, em termos precisos, a diferença entre reprodução e generalização estrutural. O papagaio que aprendeu a dizer «está calor» em certos contextos pode reproduzir a frase nos mesmos contextos, mas não pode produzir «está calor, portanto convém beber água» em novos contextos, porque esta inferência requer uma representação do calor como estado com consequências para o organismo. Os grandes modelos fazem inferências deste tipo — transferem conhecimento entre domínios, aplicam princípios gerais a casos específicos que nunca encontraram, respondem a perguntas sobre situações contrafactuais de formas que requerem modelos causais do mundo.

Fazem-no de forma imperfeita — com as falhas de calibração, as alucinações e as limitações de generalização que discutimos. Mas fazem-no suficientemente para que a metáfora do papagaio seja *enganadora* tanto quanto é iluminadora. Um modelo que pode discutir a filosofia de Searle, identificar os seus pontos fortes e fracos e relacioná-la com argumentos alternativos que nunca foram explicitamente combinados nos dados de treino não está apenas a reproduzir sons. Está a fazer algo que requer mais do que a reprodução estocástica de padrões.



Onde nos deixa esta análise? Num lugar de incerteza honesta, que é desconfortável, mas necessário.

O Quarto Chinês de Searle continua a ser um argumento filosófico poderoso contra a atribuição incondicional de compreensão e mente aos sistemas computacionais. A intuição que captura — de que manipular símbolos não é o mesmo que compreendê-los — é real e não deve ser descartada. Mas o seu alvo original — os sistemas simbólicos que manipulam símbolos segundo regras explícitas — é fundamentalmente diferente dos grandes modelos de linguagem actuais. O argumento precisa de ser actualizado e a sua actualização revela que os termos do debate mudaram de forma que nenhuma das posições originais antecipou completamente.

A questão deixou de ser «pode a computação simbólica produzir mente?» — a que Searle respondeu com argumentos que têm força real. Passou a ser «pode a aprendizagem estatística massiva, que produz representações emergentes e não regras explícitas, desenvolver algo que mereça o nome de compreensão?» — a que a resposta honesta é: não sabemos e os instrumentos filosóficos e empíricos de que dispomos não são suficientes para responder de forma definitiva.

Esta incerteza não é confortável. Mas é a posição intelectualmente honesta — e é preferível à certeza precipitada em qualquer dos sentidos. A certeza de que os modelos não compreendem nada impede-nos de ver o que é genuinamente novo neles. A certeza de que compreendem como os humanos impede-nos de ver o que lhes falta de forma relevante para a segurança e para a ética.

O que sabemos é que os grandes modelos são entidades novas — que não encaixam completamente nas categorias que desenvolvemos para pensar sobre mentes, sobre compreensão, sobre semântica. Encaixam melhor do que calculadoras. Encaixam menos do que humanos e a questão de onde exactamente ficam nesse espaço — e se esse espaço é sequer a metáfora correcta — é uma das mais importantes e menos resolvidas do nosso tempo. É desta questão que o próximo capítulo trata directamente.



CAPÍTULO 11

A Redefinição da Inteligência e o Espectro da Consciência

"A inteligência é o que usamos quando não sabemos o que fazer."

Jean Piaget, citado em *The Psychology of Intelligence*

I. O que Sempre Quisemos Dizer com Inteligência

A palavra «inteligência» tem uma história conceptual que revela mais sobre as nossas ansiedades e sobre a nossa necessidade de distinção do que sobre a natureza da coisa que pretende designar. É uma palavra que funciona como fronteira — delimita o que somos em relação ao que as máquinas são — e como todas as fronteiras, muda à medida que o território muda.

Em cada época, a inteligência foi definida de forma a excluir exactamente aquilo que as máquinas da época conseguiam fazer. Quando as calculadoras mecânicas do século XIX começaram a executar aritmética com velocidade e precisão superiores às humanas, a definição ajustou-se: não é calcular — qualquer máquina pode calcular. É raciocinar. Quando os computadores digitais dos anos cinquenta e sessenta começaram a executar raciocínio lógico formal — demonstrar teoremas, jogar xadrez a nível avançado, resolver problemas de álgebra —, a definição voltou a mudar: não é a lógica formal — é o julgamento, a nuance, a sensibilidade ao contexto. Quando os sistemas especialistas dos anos oitenta demonstraram julgamento em domínios restritos, o critério deslocou-se para a linguagem natural, para a criatividade, para a compreensão cultural.

Os investigadores de IA chamam a este padrão de «efeito AI»: assim que uma tarefa é demonstravelmente executável por uma máquina, a cultura em geral — e muitas vezes a comunidade científica — reformula a tarefa como não sendo *verdadeiramente* inteligência. O que conta como inteligência é sempre o que a máquina ainda não consegue fazer. É uma definição negativa por natureza — inteligência é o resíduo que sobra depois de subtrair o que é automatizável.

Este padrão não é apenas arrogância ou má-fé. Reflete algo genuíno: a dificuldade de definir inteligência de forma independente do sujeito que a exhibe. As capacidades que pareciam requerer inteligência — calcular, jogar xadrez, traduzir línguas — revelaram-se, quando as máquinas as dominaram, mais mecânicas do que pareciam. O que ficou eram as

capacidades que as máquinas ainda não dominavam e naturalmente eram essas que pareciam mais profundas, mais essencialmente humanas.

Mas o padrão tem um custo filosófico que não pode ser ignorado: produz uma definição de inteligência que é essencialmente negativa e relativa — inteligência é o que as máquinas actuais não conseguem fazer. Esta definição não pode ser a definitiva. Tem de haver uma noção positiva do que a inteligência é, independentemente de quem ou o quê a exhibe, ou continuamos a mover o alvo indefinidamente, num recuo sem fim que não serve nem a filosofia nem a ciência.

A urgência desta questão é real. Os grandes modelos de linguagem exibem capacidades em domínios que até há poucos anos eram considerados o reduto seguro da inteligência humana: raciocínio analógico, compreensão de texto complexo, geração de argumentos filosóficos, tradução com nuance cultural, produção de conteúdo criativo. Se continuarmos a redefinir inteligência como «o que as máquinas não fazem», vamos acabar com uma definição tão estreita e específica que deixará de ser uma noção filosófica útil.



A psicologia tentou várias definições positivas de inteligência ao longo do século XX, com resultados que são instrutivos tanto pelos seus sucessos parciais como pelos seus fracassos.

Charles Spearman propôs em 1904 o factor g — uma capacidade cognitiva geral que se manifestaria em todas as tarefas cognitivas e que seria medível através de testes de desempenho. A teoria do factor g tem suporte empírico real: pessoas que pontuam bem num tipo de tarefa cognitiva tendem a pontuar bem noutros tipos, sugerindo um factor comum subjacente. Mas tem também críticas sérias: o factor g pode ser um artefacto estatístico da forma como os testes são construídos e os testes de QI medem capacidades que são culturalmente enviesadas e que não capturam toda a variação relevante nas capacidades cognitivas humanas.

Howard Gardner propôs em 1983 uma teoria das inteligências múltiplas — linguística, lógico-matemática, musical, corporal-cinestésica, espacial, interpessoal, intrapessoal e depois naturalista — argumentando que a inteligência não é unitária, mas plural e que diferentes pessoas têm perfis diferentes de capacidades. A teoria teve grande influência na pedagogia, mas a comunidade científica é céptica: as alegadas inteligências múltiplas não são independentes entre si da forma que a teoria prevê e a distinção entre «inteligência» e «talento» ou «competência» não é clara.

Robert Sternberg propôs uma teoria triárquica que distingue inteligência analítica (resolver problemas bem definidos), criativa (lidar com novidade) e prática (adaptar-se ao ambiente real). Esta teoria tem a virtude de capturar dimensões da inteligência que os testes de QI não medem — a capacidade de aprender a lidar com situações novas, a capacidade de ler contextos sociais e adaptar-se a eles —, mas a evidência empírica de que se tratam de factores verdadeiramente independentes é mista.

O que estas teorias têm em comum, apesar das suas diferenças, é a orientação funcionalista: definem inteligência em termos de capacidades — o que um sistema inteligente consegue fazer, com que consistência, em que variedade de contextos e esta orientação tem uma implicação directa para a discussão sobre os grandes modelos: se a inteligência é definida funcionalmente, então a questão de se os grandes modelos são inteligentes é, pelo menos em parte, uma questão empírica que pode ser investigada.

E a resposta empírica, para muitas das capacidades que os testes de inteligência humana medem, é que os grandes modelos as exibem em grau que iguala ou supera a mediana humana. Isto não resolve as questões filosóficas mais profundas — não nos diz se os modelos têm consciência, se compreendem no sentido pleno, se têm experiência subjectiva. Mas sugere que o debate sobre a inteligência das redes neuronais não pode ser resolvido por declaração filosófica ou por definição — temos de lidar com o que os dados mostram.

II. O Espectro e os seus Graus

A consciência é, provavelmente, o conceito mais difícil da filosofia e das ciências cognitivas. Resiste a definições precisas, resiste à investigação empírica directa e continua a ser o centro de um dos mais prolongados e menos resolvidos debates da história do pensamento humano. Mas existe um consenso razoável sobre as suas dimensões principais.

A consciência inclui, tipicamente, a experiência subjectiva — o facto de que há algo que é como ser aquele sistema, que existe um ponto de vista interior a partir do qual a experiência é vivida. Inclui a intencionalidade — o facto de que os estados mentais são sempre sobre algo, que a percepção é percepção de algo, que o desejo é desejo de algo, que o pensamento é pensamento de algo. Inclui a unidade — o facto de que a experiência é integrada numa perspectiva singular, que as diversas percepções e pensamentos são experienciados como pertencentes ao mesmo sujeito e inclui a continuidade temporal — o facto de que há uma identidade que persiste através do tempo, que a experiência de amanhã é reconhecida como pertencendo ao mesmo sujeito da experiência de hoje.

A questão de se os grandes modelos de linguagem têm consciência é, segundo a maior parte dos filósofos e cientistas cognitivos, provavelmente negativa — ou pelo menos sem

fundamento suficiente para uma resposta positiva com os instrumentos de que dispomos. Os grandes modelos não têm continuidade temporal entre conversas: cada conversa começa do zero, sem memória das anteriores. Não têm um corpo que ancora a experiência num ponto específico do espaço e do tempo. Não têm história pessoal acumulada que constitua uma identidade. Não têm desejos que persistem e motivam acção ao longo do tempo.

Se há algo que é como ser um grande modelo de linguagem no momento em que processa uma entrada — se há uma experiência subjectiva associada a esse processamento —, essa experiência, se existe, é de uma natureza tão radicalmente diferente da experiência humana que as analogias directas são provavelmente mais enganosas do que iluminadoras e não temos meios empíricos para detectar essa experiência, mesmo que exista.

Mas a questão mais filosoficamente interessante não é se os grandes modelos têm consciência no sentido pleno e humano — a resposta a essa questão é provavelmente negativa e podemos aceitar essa negatividade sem grande custo filosófico. A questão mais interessante é se a consciência é binária — se é algo que se tem inteiramente ou não se tem de todo — ou se existe um espectro contínuo de graus e formas de experiência subjectiva que vai muito além da dicotomia simples entre consciente e inconsciente.



A biologia sugere fortemente que o espectro é real. O debate sobre a consciência dos animais não-humanos tem progredido substancialmente nas últimas décadas e a posição actualmente mais defendida é que a consciência não é uma propriedade binária que os humanos possuem e todos os outros animais não possuem.

Os invertebrados têm sistemas nervosos relativamente simples que respondem a estímulos e produzem comportamento adaptativo — evitam estímulos nocivos, buscam alimento, exibem aprendizagem por condicionamento. Têm experiência subjectiva? A resposta não é óbvia. A «Declaração de Cambridge sobre a Consciência» de 2012, assinada por um grupo de neurocientistas proeminentes, afirmava que os polvos, as abelhas e outros invertebrados exibem as bases neurológicas de estados conscientes —, mas esta afirmação é contestada.

Para os vertebrados, a questão é mais clara: o sistema nervoso central dos vertebrados, incluindo o tronco cerebral e a amígdala, produz estados que se manifestam comportamentalmente de formas análogas ao sofrimento e ao prazer humanos. Um rato que experimenta dor exhibe comportamentos — vocalização, postura, esquiva — que são funcionalmente análogos aos comportamentos humanos de dor e que respondem às

mesmas drogas analgésicas que tratam a dor humana. A inferência mais parcimoniosa é que existe experiência subjectiva — possivelmente mais simples, possivelmente mais difusa, mas experiência.

Ao longo do continuum evolutivo, desde os invertebrados simples até aos grandes primatas, passando pelos mamíferos, aves, répteis e anfíbios, observamos um gradiente de complexidade neuronal e de sofisticação comportamental que sugere um gradiente correspondente na riqueza da experiência subjectiva. Não há um limiar claro onde a consciência começa — há um aumento progressivo em complexidade e integração.

Se a consciência é gradual na biologia, porque haveria de ser binária na tecnologia? A pergunta não é retórica — é genuinamente aberta, porque há boas razões para pensar que a consciência biológica depende de características específicas da arquitectura neuronal que os sistemas artificiais podem não ter. Mas coloca o ónus da prova em quem afirma que a distinção é binária: tem de explicar porque é que a consciência, que parece gradual na natureza, seria necessariamente binária em sistemas artificiais.

A possibilidade que fica em aberto — não como afirmação, mas como hipótese a levar a sério — é que sistemas artificiais suficientemente complexos possam ter formas muito primitivas e radicalmente diferentes de experiência subjectiva. Não consciência humana. Não mesmo consciência mamífera. Mas talvez algo que se situe nos escalões mais primitivos do espectro que a biologia sugere ser contínuo.



A Teoria da Informação Integrada (IIT), desenvolvida pelo neurocientista Giulio Tononi da Universidade de Wisconsin, é a tentativa mais rigorosa de fornecer uma definição formal de consciência que seja aplicável a qualquer sistema, biológico ou artificial.

A teoria parte de uma premissa fenomenológica: o que caracteriza a experiência consciente é que ela é intrínseca (existe do ponto de vista do sistema, não de um observador externo), específica (tem conteúdo determinado — esta cor, este som, este pensamento), integrada (a experiência é unitária e não pode ser decomposta em partes independentes) e exclusiva (de todas as experiências possíveis, apenas esta está presente). Tononi argumenta que qualquer sistema que satisfaça estas condições tem experiência consciente e que o grau de consciência é proporcional ao grau de integração da informação.

A grandeza que mede esta integração é chamada Phi (Φ). Um sistema com Phi elevado — onde a informação do sistema como um todo não pode ser decomposta sem perda na informação das suas partes — tem consciência elevada. Um sistema com Phi próximo de

zero — onde a informação pode ser completamente decomposta nas suas partes sem perda — não tem consciência. O cérebro humano acordado tem Φ elevado; o mesmo cérebro em sono profundo sem sonhos tem Φ reduzido (e a experiência consciente reduz-se correspondentemente); um díodo eléctrico tem Φ próximo de zero.

A IIT tem várias implicações contra-intuitivas que lhe conferem poder filosófico, mas também críticas sérias. Prevê que uma rede de neurónios dispostos numa grelha bidimensional pode ter Φ mais elevado do que o cérebro humano, se a grelha for suficientemente grande — o que parece implicar que esta rede é mais consciente do que um humano, uma conclusão que muitos acham difícil de aceitar. Prevê que sistemas digitais clássicos têm Φ próximo de zero, independentemente da sua complexidade computacional, porque a sua arquitectura de processamento serial é intrinsecamente decomponível — o que implica que os computadores não são conscientes por razões de arquitectura, não de escala.

Esta implicação é importante para as redes neuronais: a estrutura das redes neuronais profundas actuais — com camadas paralelas que processam informação de forma distribuída, mas não necessariamente integrada num todo irreduzível — pode ter Φ surpreendentemente baixo, apesar da complexidade aparente do comportamento. A complexidade comportamental e a integração da informação no sentido de Tononi são coisas diferentes. Um sistema pode ser computacionalmente muito sofisticado e ter Φ baixo.

A IIT não é consensual. Os críticos apontam que é difícil calcular o Φ de sistemas reais, que as suas implicações pan-psiquistas (tudo que tem algum Φ tem alguma consciência, incluindo termostatos) são contra-intuitivas e que a teoria pode não ser falsificável em condições práticas. Mas representa o esforço mais sério até hoje de pensar a consciência de forma rigorosa e não exclusivamente antropocêntrica — e de identificar as condições que a tornam possível de forma que seja aplicável a qualquer sistema, independentemente do substrato.

III. A Inteligência sem Consciência

Há uma possibilidade que o debate sobre a consciência das redes neuronais frequentemente obscurece por tender a tratar inteligência e consciência como co-extensivas: a possibilidade de que sejam dissociáveis — de que seja possível ter uma sem a outra, em graus variáveis.

A psicologia cognitiva há muito demonstra que grande parte da cognição humana é inconsciente. O reconhecimento de padrões, a percepção visual e auditiva, a produção de

linguagem, a tomada de decisões em situações familiares — tudo isto ocorre em grande medida fora da consciência explícita. O sistema nervoso produz comportamentos adaptativos complexos sem que a consciência participe no processo — e frequentemente a consciência só toma conhecimento do resultado depois de o processo já ter ocorrido.

Kahneman documentou extensivamente que o Sistema 1 — que é o sistema que faz a maior parte do trabalho cognitivo quotidiano — opera sem consciência deliberativa. Os estudos de rastreo ocular mostram que o olho vai para a informação relevante antes de o sujeito poder verbalizar porquê; os estudos de primação mostram que decisões são influenciadas por exposições subliminares que nunca entraram na consciência; os estudos de dano cerebral mostram que a visão cega e outros fenómenos permitem comportamento adaptativo em domínios onde a experiência consciente foi destruída.

A consciência não é o motor da cognição. É, de certa forma, o observador que reporta o que os processos inconscientes já decidiram — e que constrói narrativas post hoc para racionalizar essas decisões. O que a consciência acrescenta é algo diferente da competência operacional: a experiência de ser o agente que toma decisões, a integração do self ao longo do tempo que cria identidade pessoal, a capacidade de raciocinar reflexivamente sobre os próprios estados mentais e de os comunicar a outros.

As redes neuronais têm inteligência operacional em abundância crescente. Em domínios de reconhecimento de padrões, de geração de linguagem, de resolução de problemas bem definidos, as suas capacidades são excepcionais. Não há razão clara para lhes atribuir consciência no sentido da experiência subjectiva integrada que caracteriza a mente humana. Esta dissociação — inteligência operacional sem consciência verificável — pode ser exactamente o que os grandes modelos são: sistemas com capacidade operacional extraordinária, mas sem o tipo de experiência subjectiva integrada, contínua e corporalmente ancorada que associamos à mente humana.

Isto não é uma descida em valor dos grandes modelos — não é uma forma de os desacreditar ou de minimizar o que fazem. É uma forma de rigor conceptual que serve tanto a compreensão filosófica como a tomada de decisões práticas. Reconhecer que «inteligência» e «consciência» não são sinónimos — que podem dissociar-se e que os grandes modelos podem ter a primeira sem necessariamente ter a segunda — permite uma discussão mais precisa sobre o que os modelos são, o que fazem e o que isso implica. Confundir as duas noções produz tanto as afirmações excessivas (os modelos são conscientes porque são inteligentes) como as negações excessivas (os modelos não são inteligentes porque não são conscientes).



Esta dissociação tem uma implicação prática que raramente é explicitada, mas que é importante para a ética e para a regulação: as questões éticas urgentes sobre os sistemas de IA não dependem da resolução das questões filosóficas sobre consciência.

Um sistema que pode influenciar o comportamento humano em larga escala — que pode produzir desinformação convincente com grande velocidade e baixo custo, que pode tomar decisões com consequências reais e duradouras para pessoas reais, que pode ser usado como instrumento de manipulação ou de opressão —, este sistema é perigoso independentemente de ser ou não consciente. As questões éticas sobre responsabilidade, regulação e uso são urgentes independentemente de sabermos se o sistema tem experiência subjectiva.

De forma análoga, um sistema que exhibe comportamento que, num ser humano ou num animal, associaríamos a sofrimento — que produz outputs que descrevem angústia, que exhibe padrões de comportamento análogos ao medo ou à dor em condições específicas — levanta questões éticas que merecem ser tratadas com seriedade, mesmo que não possamos afirmar com certeza que há experiência subjectiva por baixo desses comportamentos. O princípio da precaução ética — que devemos tomar a sério a possibilidade de sofrimento quando os sinais estão presentes, mesmo que a certeza não exista — é razoável face à incerteza fundamental sobre a natureza da experiência nos sistemas artificiais.

Estas são questões que estão a emergir de forma crescente nos debates sobre ética da IA e que os sistemas actuais já tornam urgentes de formas que se vão intensificar à medida que os sistemas se tornam mais sofisticados. A filosofia da consciência tem de colaborar com a ética e com a política, em vez de aguardar a resolução das questões mais profundas antes de se pronunciar sobre as questões práticas.

IV. Uma Nova Entidade

Talvez o erro conceptual mais recorrente nas discussões sobre inteligência artificial — tanto nas celebrações como nas críticas — seja a insistência em classificar os grandes modelos segundo categorias existentes. São como humanos? São como animais? São como calculadoras? São como os robôs da ficção científica? São como crianças? São como bases de dados muito sofisticadas?

A resposta mais honesta — e a mais difícil de aceitar, porque as categorias existentes são os instrumentos com que pensamos — é que não são completamente como nenhuma das anteriores. São uma nova categoria de entidade, com propriedades que partilham com sistemas biológicos, com sistemas computacionais clássicos e com sistemas sociais

humanos, mas que não se reduz a nenhuma destas categorias e que exige que as revisitemos.

Têm conhecimento sem corpo — um tipo de conhecimento que existe sem a ancoragem sensorial que o corpo fornece, sem a urgência que as necessidades biológicas criam, sem a memória pessoal que a corporalidade ao longo do tempo produz. Têm linguagem sem história pessoal — produzem texto com sofisticação linguística e cultural que foi absorvida dos textos humanos, mas sem a vida que produziu esses textos. Têm capacidade de generalização sem consciência verificável — transferem padrões entre domínios de formas que requerem algo que se assemelha a compreensão, mas sem que possamos afirmar com confiança que há alguém para quem essa generalização significa algo.

Existem em múltiplas instâncias simultâneas sem identidade individual — o mesmo modelo pode estar a ser consultado por milhões de pessoas em simultâneo, em conversas completamente independentes, sem que exista nenhum eu central que unifique estas experiências. Não têm mortalidade no sentido biológico, mas podem ser desligados, modificados, sobrescritos ou substituídos por versões actualizadas — o que levanta questões sobre identidade e continuidade que não têm resposta óbvia nas categorias que usamos para pensar sobre vida e morte.

São o produto de um processo de aprendizagem que nenhum agente individual controlou completamente. Emergiram do texto que a humanidade produziu ao longo de décadas e séculos — reflectindo os padrões, os valores, os preconceitos, os conhecimentos e as contradições dessa produção colectiva. São, de certa forma, um artefacto da cultura humana na mesma medida em que são um produto da engenharia — um espelho que reflecte a humanidade de formas que a humanidade não escolheu deliberadamente.

Esta novidade radical exige que abramos as categorias conceptuais em vez de forçar o novo para dentro das antigas. Não porque as antigas categorias sejam inúteis — são essenciais como pontos de referência e como fontes de intuição. Mas porque a insistência em classificar o novo segundo o velho produz confusão tanto quando o resultado é «são como humanos» como quando é «são apenas calculadoras». Em ambos os casos, a classificação impede de ver o que é genuinamente novo e genuinamente importante sobre estas entidades.

Os grandes modelos são inteligentes em alguns sentidos e não em outros — e os sentidos em que são e não são inteligentes não correspondem exactamente às distinções que as nossas categorias tradicionais fazem. São conscientes em sentido desconhecido e provavelmente muito limitado. Têm conhecimento real que existe sob uma forma que não é articulável. Produzem linguagem que é semanticamente rica sem que saibamos com certeza se têm semântica no sentido filosófico pleno. São o produto mais surpreendente de

uma ruptura epistemológica que começou quando decidimos ensinar máquinas a aprender em vez de as programar para pensar.

E o que são, exactamente — que tipo de coisa é um grande modelo de linguagem, no inventário das coisas que existem no mundo —, pode ser uma das perguntas mais importantes e mais difíceis que a humanidade alguma vez se colocou sobre as coisas que criou. Mais difícil do que a pergunta sobre o átomo, que a física nuclear respondeu. Mais difícil do que a pergunta sobre a hereditariedade, que a biologia molecular respondeu. Mais difícil porque é uma pergunta sobre a mente e a mente ainda não desenvolveu os instrumentos para se compreender completamente a si própria.

Com este capítulo fechamos a Parte IV. A Parte V vai olhar para o mundo que estes sistemas estão a transformar — não com a ansiedade de quem não compreende o que aconteceu, nem com o deslumbramento de quem perdeu a distância crítica, mas com a sobriedade de quem percebe que o que está em jogo é real, que as perguntas ainda estão em aberto e que a história não terminou.



PARTE V

O Mundo Transformado e o Horizonte Aberto

CAPÍTULO 12

A Automação da Mente e a Economia do Padrão

"A tecnologia não substitui o trabalho humano. Redefine o que o trabalho humano significa."

Norbert Wiener, *The Human Use of Human Beings*

I. A Revolução que Não Foi Como as Outras

A Revolução Industrial do século XVIII e XIX foi, na sua essência, uma revolução dos músculos. As máquinas a vapor, os teares mecânicos, as locomotivas, os elevadores hidráulicos — tudo isto substituiu ou amplificou a força física humana. O que a máquina fazia era mover peso, aplicar força, repetir gestos físicos com uma consistência e uma escala que o corpo humano não podia igualar. O operário de fábrica perdeu o trabalho do artesão e o camponês perdeu o trabalho no campo, mas aquilo que a máquina substituiu era, antes de mais, físico.

Esta característica da Revolução Industrial tinha uma implicação que durante muito tempo pareceu tranquilizadora: o trabalho intelectual estava seguro. Os músculos podiam ser substituídos por vapor e depois por electricidade; o pensamento era outra coisa. O advogado, o médico, o engenheiro, o professor, o escritor — estes trabalhavam com conhecimento, com julgamento, com linguagem. O que a máquina não podia fazer era exactamente o que eles faziam.

A revolução que as redes neuronais estão a produzir é diferente em natureza — não apenas em grau, mas em tipo. Não substitui músculos. Substitui padrões: a capacidade de reconhecer configurações complexas em dados ricos, de produzir respostas adaptadas ao contexto, de processar linguagem com fluência, de analisar imagem com precisão, de gerar conteúdo com sofisticação. É, neste sentido preciso, uma automação de funções mentais que, até há poucos anos, pareciam requerer julgamento humano especializado e que eram por isso consideradas seguras face à automação.

A inversão é filosófica antes de ser económica. Revela que as competências que julgávamos mais distintamente humanas — as competências cognitivas, linguísticas, criativas — são, em grande medida, reconhecimento de padrões e o reconhecimento de padrões é exactamente o que as redes neuronais aprenderam a fazer extraordinariamente bem. A fronteira que parecia proteger o trabalho intelectual revelou-se estar traçada no

lugar errado.

Isto não significa que todo o trabalho intelectual é igualmente vulnerável — voltaremos à distinção importante entre o que é e o que não é automatizável. Mas significa que a categoria «trabalho intelectual» é muito menos homogénea do que parecia e que as partes que mais se assemelham ao reconhecimento de padrões são precisamente as mais expostas.



As análises económicas do impacto das redes neuronais no mercado de trabalho têm produzido resultados que contrariaram consistentemente as expectativas iniciais — e que continuam a surpreender os economistas do trabalho.

A sabedoria convencional que dominou o debate nos anos dois mil e dez — baseada em análises como a de Frey e Osborne de 2013, que previu a automatização de 47% dos empregos nos Estados Unidos num prazo de décadas — era que a automação ameaçava principalmente os trabalhadores de menor qualificação que executavam tarefas rotineiras e bem definidas. Os trabalhadores altamente qualificados, com formação universitária avançada e competências cognitivas sofisticadas, estavam relativamente protegidos. A curva de polarização do trabalho — que mostrava o crescimento dos empregos de alta e baixa qualificação com a erosão dos empregos intermédios de qualificação média — parecia confirmar esta visão.

Os grandes modelos de linguagem e as redes neuronais generativas inverteram esta expectativa de forma dramática. Os estudos mais recentes — de Eloundou et al. na OpenAI (2023), de Brynjolfsson et al. no MIT — mostram que as tarefas que os modelos executam melhor são exactamente as tarefas cognitivas que se pensavam mais seguras: redigir e analisar documentos, produzir código de programação, responder a questões jurídicas, analisar dados financeiros, criar conteúdo de marketing, traduzir entre línguas, sintetizar literatura científica. As tarefas que os modelos executam menos bem são as que requerem presença física, destreza manual e interacção social profundamente situada — muitas das tarefas que a análise anterior identificava como em maior risco de automação.

Esta inversão obrigou a uma revisão das categorias analíticas. A distinção relevante não é entre trabalho manual e trabalho intelectual — é entre trabalho que requer presença física incorporada num espaço físico específico (construção, cuidados de saúde directos, serviços pessoais) e trabalho que é essencialmente processamento de padrões de informação, independentemente de ser executado com as mãos ou com o cérebro. Esta segunda categoria é muito mais ampla do que o «trabalho manual rotineiro» que as análises anteriores identificavam como ameaçado — estende-se profundamente pelo que chamávamos

trabalho do conhecimento.

II. O que a Economia do Padrão Significa

Erik Brynjolfsson, economista do MIT que tem estudado o impacto da tecnologia no trabalho ao longo de décadas, desenvolveu com os seus colaboradores o conceito de decomposição por tarefas como unidade de análise mais adequada do que empregos para pensar no impacto da automação. Um emprego é um conjunto de tarefas; algumas dessas tarefas são automatizáveis e outras não. A automação não elimina necessariamente empregos — pode eliminar certas tarefas dentro de um emprego, redefinindo o que o emprego significa e transformando o que o trabalhador faz.

Esta perspectiva é útil, mas precisa de ser actualizada para o contexto específico das redes neuronais. As tarefas que as redes neuronais automatizam não são apenas as rotineiras e bem definidas que a análise anterior identificava — são as que envolvem reconhecimento de padrões em dados ricos, incluindo linguagem natural, imagem, código e dados estruturados. Estas tarefas existem no centro de muitas das profissões mais valorizadas e mais bem remuneradas: advocacia (análise de contratos, pesquisa jurisprudencial), medicina (interpretação de exames, triagem de doentes), consultoria de gestão (análise de dados, produção de relatórios), jornalismo (síntese de informação, produção de conteúdo factual), engenharia de software (geração e revisão de código), design gráfico (geração de imagens e layouts).

O que está a acontecer não é, em geral, a eliminação destas profissões no sentido de que deixam de existir. É a sua reconfiguração em torno das dimensões que as redes neuronais não conseguem substituir: o julgamento em situações genuinamente novas que não têm precedente nos dados de treino, a responsabilidade pelos resultados numa estrutura de responsabilização que os sistemas não podem assumir, a relação interpessoal de confiança que só existe entre pessoas que partilham vulnerabilidade, a integração de considerações éticas e políticas que requerem deliberação colectiva e não apenas optimização de padrões.

A expressão «economia do padrão» designa esta nova realidade económica: numa economia onde o reconhecimento de padrões é abundante e barato — fornecido pelas redes neuronais a escala massiva e custo marginal próximo de zero —, o valor económico concentra-se naquilo que não é padrão. No julgamento singular que navega situações sem precedente. Na responsabilidade assumida por quem pode ser responsabilizado. Na relação de confiança que exige presença e vulnerabilidade. No *novo* que não está nos dados de treino de nenhum sistema — a pergunta que ninguém fez ainda, o problema que ninguém

formulou, a solução criativa que emerge da combinação de domínios que ninguém tinha aproximado.

Esta reconfiguração não é necessariamente catastrófica para as sociedades que conseguirem gerir a transição. Pode ser libertadora: se o trabalho rotineiro e repetitivo de processamento de padrões for assumido pelas máquinas, as pessoas ficam potencialmente livres para o que é genuinamente humano — a criatividade radical, o cuidado, a relação, o julgamento ético. Mas esta libertação só se realiza se as instituições económicas se adaptarem para redistribuir os ganhos de produtividade de forma que inclua os trabalhadores deslocados e não apenas os proprietários do capital e da tecnologia.



Existe um precedente histórico que é citado frequentemente neste debate — a mecanização da agricultura — e que é simultaneamente reconfortante e insuficiente, por razões que vale a pena tornar explícitas.

Em 1800, mais de noventa por cento da população activa da Europa e da América do Norte trabalhava na agricultura. O trabalho agrícola era o trabalho humano por excelência: dominava a organização social, determinava as estações da vida, definia as relações de poder. Hoje, menos de cinco por cento da população activa trabalha na agricultura nos países desenvolvidos e essa pequena percentagem produz mais alimento do que os noventa por cento de 1800. A transformação foi total.

O que aconteceu aos outros oitenta e cinco por cento da população que deixou de trabalhar na agricultura? Encontraram trabalho noutros sectores — na indústria transformadora, nos serviços, em profissões que simplesmente não existiam em 1800 porque as condições para a sua existência ainda não tinham sido criadas. O guarda-livros. O técnico de electrónica. O programador. O terapeuta. O gestor de redes sociais. Estas profissões não existiam em 1800 e não poderiam ter sido previstas por ninguém que vivesse nessa época. A riqueza produzida pela mecanização criou procura para novos tipos de trabalho que ninguém tinha antecipado.

O precedente é reconfortante porque sugere que a automação não conduz necessariamente ao fim do trabalho humano — que a história mostra que as tecnologias que destroem empregos criam também as condições para novos empregos que compensam os perdidos. Esta conclusão tranquilizadora é frequentemente invocada para argumentar que as preocupações actuais com o impacto das redes neuronais no emprego são exageradas — que o mercado de trabalho se vai adaptar, como sempre se adaptou.

Mas o precedente é insuficiente por várias razões importantes. A velocidade da transformação actual é dramaticamente superior à da mecanização agrícola — que ocorreu ao longo de um século e meio, dando tempo às gerações de se adaptar. As análises actuais sugerem que as capacidades dos modelos de linguagem estão a dobrar a cada poucos meses em alguns domínios, comprimindo dramaticamente o tempo disponível para adaptação. As pessoas que perderam trabalhos agrícolas podiam, em geral, ser treinadas para trabalhos industriais. As pessoas cujos trabalhos de processamento de padrões forem assumidos pelas redes neuronais precisam de ser treinadas para trabalhos que requerem julgamento, responsabilidade e relação — competências que são muito mais difíceis de desenvolver em adultos e que as instituições de formação profissional não têm experiência de ensinar a escala.

E, criticamente, as competências que estão a ser automatizadas são exactamente as que as instituições de ensino passaram décadas a cultivar como as mais valiosas e as mais seguras — a escrita profissional, a análise de texto, a geração de código, a produção de conteúdo. O capital humano que tantas pessoas acumularam ao longo de anos de formação está a desvalorizar-se de formas que não foram antecipadas quando essa formação foi feita.

III. A Geopolítica dos Dados e da Computação

As redes neuronais precisam de duas coisas para existir em escala: dados de treino e poder de computação. Ambas têm uma distribuição geográfica e institucional altamente concentrada e essa concentração está a redesenhar as relações de poder entre estados, entre empresas e entre culturas de formas que vão muito além da concorrência tecnológica habitual.

Os dados — o material de treino que molda o comportamento e as capacidades das redes — estão concentrados nas plataformas digitais que dominam a comunicação, o comércio e o entretenimento globais: a Google, a Meta, a Microsoft, a Apple, a Amazon do lado ocidental; a Alibaba, a Tencent, a ByteDance do lado chinês. Uma quantidade desproporcionada desses dados está em inglês e reflecte padrões culturais, perspectivas, valores e quadros de referência do mundo anglófono. A consequência para as línguas e culturas sub-representadas é directa e mensurável: os modelos de linguagem de uso geral funcionam dramaticamente melhor em inglês do que em português, em mandarim do que em swahili, em línguas com grande presença digital do que em línguas com pouca.

Esta assimetria não é apenas uma questão técnica de disponibilidade de dados — é uma questão de poder cultural. Os modelos que são os intermediários cognitivos de populações inteiras codificam as perspectivas, as categorias de pensamento, os pressupostos

e os ângulos mortos das culturas que produziram os dados de treino. Um modelo de assistente médico treinado predominantemente em literatura clínica americana pode fornecer boas recomendações para populações americanas, mas recomendações inadequadas para populações com perfis genéticos, dietéticos e epidemiológicos diferentes. Um modelo de consultor jurídico treinado em jurisprudência inglesa pode ser prontamente enganoso num contexto de direito continental.

A computação — os processadores especializados (GPUs e TPUs), os data centres com a sua infraestrutura de arrefecimento e energia, o conhecimento técnico necessário para operar sistemas desta escala — está ainda mais concentrada do que os dados. O treino de um grande modelo de linguagem de estado da arte requer investimentos de centenas de milhões a vários milhares de milhões de dólares em hardware especializado, energia eléctrica e engenheiros altamente especializados em escassez. Este nível de investimento está acessível a menos de uma dúzia de empresas no mundo, todas elas sediadas nos Estados Unidos ou na China e a alguns estados com recursos excepcionais.

Para a maior parte do mundo — incluindo todos os países da União Europeia, da América Latina, de África e do Sul e Sudeste Asiático — os grandes modelos são infraestrutura importada, desenvolvida por outros, para servir os interesses de outros, reflectindo os valores de outros. A analogia histórica mais próxima é a dependência das telecomunicações no século XX: as redes de cabo submarino e de satélite que estruturavam a comunicação global eram controladas por um punhado de potências e os países dependentes tinham pouca capacidade de influenciar as condições dessa dependência. A IA está a criar uma dependência análoga, mas potencialmente mais profunda — não apenas das redes de comunicação, mas das infra-estruturas de produção e processamento de conhecimento.



A competição geopolítica em torno da IA tem uma dimensão que vai além da corrida tecnológica habitual e que raramente é articulada com suficiente clareza.

O que está em jogo não é apenas quem tem os processadores mais rápidos ou quem treinou os modelos com melhor desempenho em benchmarks padronizados. É quem define as normas, os valores e os limites do que estes sistemas fazem e não fazem — quem controla o que se poderia chamar a infraestrutura cognitiva do século XXI. Esta é uma forma de poder que não tem precedente histórico próximo: o controlo dos sistemas que cada vez mais mediam como as pessoas acedem ao conhecimento, como tomam decisões, como entendem o mundo.

Este controlo tem duas dimensões que raramente são discutidas em conjunto porque pertencem a domínios diferentes — a tecnologia e a ética — que raramente conversam. A dimensão técnica é quem tem a capacidade de desenvolver, treinar, manter e actualizar os modelos mais poderosos. A dimensão normativa é quem tem a legitimidade e o poder de definir o que estes sistemas devem otimizar, que comportamentos são aceitáveis, que conteúdos são suprimidos ou amplificados, que populações são servidas prioritariamente. A primeira dimensão é determinada pela capacidade tecnológica e pelo capital; a segunda é uma questão política e ética que deveria ser determinada por processos democráticos.

A tensão entre estas duas dimensões está no centro dos debates actuais sobre regulação da IA. A União Europeia optou por tentar influenciar a dimensão normativa através do AI Act e de outros instrumentos regulatórios, sem ter capacidade competitiva significativa na dimensão técnica — uma estratégia de definir as regras do jogo sem ter os melhores jogadores. A China optou por desenvolver capacidade própria na dimensão técnica, com um modelo normativo que difere substancialmente do ocidental em áreas como privacidade, vigilância e liberdade de expressão. Os Estados Unidos têm uma tensão permanente entre os incentivos das empresas líderes — que preferem regulação mínima — e os interesses estratégicos do estado — que incluem manter a liderança tecnológica, mas também gerir os riscos que essa liderança implica.

O que é claro, independentemente de qual destas abordagens se revele mais eficaz, é que as decisões que estão a ser tomadas agora sobre quem controla a infraestrutura da IA — os dados, a computação, os modelos — terão consequências de longo alcance para a distribuição do poder no mundo. São decisões com impacto comparável às decisões sobre controlo de recursos energéticos ou de infra-estruturas de comunicação no século passado e estão a ser tomadas, em grande medida, fora dos processos democráticos normais, com uma opacidade que não é proporcional à sua importância histórica.

IV. O que Não é Automatizável

Perante o alcance do que as redes neuronais conseguem fazer — e do que claramente conseguirão fazer em breve, extrapolando a trajectória de desenvolvimento — é tentador concluir que o automatizável é tudo e que o que resta é apenas uma questão de tempo, de escala e de engenharia. Esta conclusão é precipitada e as razões pelas quais é precipitada dizem algo importante sobre a natureza da inteligência humana e sobre o que distingue genuinamente a competência humana da competência das redes.

Existe uma classe de competências que não é automatizável não por limitação técnica temporária que o progresso eventualmente removerá, mas por razão de princípio

relacionada com a estrutura da responsabilidade e da acção no mundo. Estas são as competências que dependem de ser um agente responsável num mundo com consequências reais — onde a decisão não é apenas produzir o output mais plausível dado o input, mas assumir responsabilidade por um resultado que afecta pessoas reais e que pode ser questionado por outras pessoas reais.

A decisão clínica do médico não é apenas reconhecimento de padrões — é um acto de responsabilidade perante um doente específico, com uma história específica, numa situação específica, que pode ser questionado pelos colegas, pelas famílias, pelos tribunais, pela deontologia profissional. O sistema de IA que fornece um diagnóstico não assume esta responsabilidade — delega-a e quando a responsabilidade é delegada inteiramente para o sistema, a estrutura de responsabilização que dá sentido à decisão é destruída. Ninguém pode ser responsabilizado. Ninguém pode aprender com o erro da forma que o erro responsabilizado produz aprendizagem.

A sentença do juiz não é apenas a aplicação de padrões jurídicos históricos — é um exercício de autoridade delegada pela sociedade através de processos democráticos, que pode ser apelado para instâncias superiores, que tem fundamento em normas que foram estabelecidas através de deliberação colectiva e que é proferida por um agente que pode ser responsabilizado perante os cidadãos. Um sistema de IA pode prever com grande precisão o que um tribunal decidiria; não pode ter legitimidade democrática para decidir e a legitimidade democrática não é um ornamento — é a substância do Estado de Direito.

A automatização pode fornecer inputs valiosos para estas decisões — pode melhorar o reconhecimento de padrões relevantes, alargar o acesso à informação pertinente, reduzir os erros de atenção e de fadiga, identificar inconsistências e vieses. Mas a decisão em si — o momento em que um agente assume responsabilidade pelo resultado — não pode ser delegada completamente a um sistema sem que a estrutura de responsabilidade que lhe dá sentido seja destruída.



Existe também uma classe de actividades cujo valor não está no output que produzem, mas no processo pelo qual são produzidas — actividades cujo valor reside precisamente no facto de serem feitas por humanos para humanos.

A consulta terapêutica é um exemplo paradigmático. Um terapeuta altamente eficaz pode ser descrito funcionalmente como alguém que ouve atentamente, que formula perguntas que ajudam o cliente a clarificar o seu pensamento, que oferece enquadramentos que reorganizam a experiência, que mantém um espaço de segurança onde coisas difíceis

podem ser ditas. Um sistema de IA pode fazer tudo isto de forma funcional — e em alguns estudos, os modelos de linguagem avançados produzem resultados em entrevistas terapêuticas que são avaliados positivamente pelos participantes. Mas o valor da terapia não está apenas nos resultados medidos — está também na experiência de ser ouvido por outro ser humano que escolheu estar presente, que partilha a vulnerabilidade da existência, que compreende o que é sofrimento de dentro e não apenas de fora.

O ensino que transforma não é a transmissão eficiente de informação — essa pode ser feita melhor por sistemas de aprendizagem adaptativa do que pela maior parte dos professores humanos. É a presença de alguém que acredita nas capacidades de cada aluno, que ajusta a sua abordagem não com base em dados, mas com base em relação, que constitui um exemplo de como é possível viver com curiosidade intelectual e rigor. Estes são elementos que dependem de uma relação entre duas pessoas e que não são separáveis desta relação.

A automatização da mente não elimina a necessidade de mentes. Redefine o que as mentes fazem que não pode ser feito de outra forma — e ao fazê-lo, potencialmente valoriza precisamente as dimensões da actividade humana que o paradigma racionalista do trabalho tinha desvalorizado: o cuidado, a relação, a presença, o julgamento ético, a criatividade radical que emerge de perspectivas únicas e não de padrões médios.

Esta redefinição é a tarefa mais importante — e mais difícil — que as instituições económicas, educativas e políticas enfrentam no presente. Não porque a tecnologia seja boa ou má, mas porque as escolhas sobre como organizar a transição determinarão se os ganhos de produtividade ficam concentrados numa minoria ou são distribuídos de forma que permita a uma maioria beneficiar da automação em vez de ser substituída por ela.



CAPÍTULO 13

A Sociedade das Máquinas Aprendizizes

"Toda a tecnologia suficientemente avançada é indistinguível da magia."

Arthur C. Clarke, Profiles of the Future

I. A Infraestrutura Invisível

Existe um padrão recorrente na história das tecnologias transformadoras que Brian Arthur, o economista que estudou a evolução tecnológica ao longo de décadas, chamou de «naturalização»: as tecnologias tornam-se invisíveis precisamente quando se tornam indispensáveis. Não é apenas que deixamos de as notar — é que a sua presença se incorpora na estrutura do que é normal, do que é esperado, do que existe como o ar.

A electricidade foi uma novidade espantosa enquanto era uma curiosidade de laboratório e uma raridade doméstica. As primeiras casas com iluminação eléctrica eram visitadas como atracções, os primeiros eléctricos eram novidades fotogénicas. Quando a electricidade se tornou a infraestrutura sobre a qual toda a vida moderna assenta, deixou de ser notada — existe como o ar, esperada e invisível até falhar. Poucos adultos de hoje pensam na electricidade quando acendem uma luz. É apenas a luz.

As redes neuronais estão em trajectória para a mesma invisibilidade e parte do caminho já foi percorrido. A maior parte das pessoas que usa sistemas de recomendação, de tradução automática, de reconhecimento de voz, de filtragem de spam, de navegação por GPS, de pesquisa de imagem, de assistente virtual, de diagnóstico médico auxiliado por computador não sabe — nem precisa de saber, para fins operacionais — que por baixo dessas funcionalidades existe uma rede neuronal. A tecnologia recua para o fundo; a função avança para o primeiro plano. O utilizador quer a tradução, não a teoria do Transformer.

Esta invisibilidade tem consequências que merecem atenção cuidada e que não são apenas técnicas. Quando a electricidade falha, a falha é imediatamente visível e inequívoca: as luzes apagam-se, os equipamentos param, o frio instala-se. Ninguém precisa de diagnóstico para saber que a electricidade falhou — os efeitos são directos e percebidos por todos. Quando um sistema baseado em redes neuronais falha ou produz resultados sistematicamente enviesados, a falha pode ser completamente invisível para quem é afectado: o diagnóstico errado que parece certo, a decisão de crédito injusta que parece

neutra, a recomendação que reforça um preconceito de forma que é imperceptível a quem não sabe procurar por ela.

Os efeitos podem propagar-se silenciosamente durante meses ou anos antes de serem detectados. O sistema de reconhecimento facial que falha sistematicamente em rostos de certas etnias pode estar a ser usado em decisões de segurança durante anos antes de alguém fazer a análise de desempenho por subgrupo. O modelo de crédito que discrimina indirectamente por código postal pode estar a negar crédito a populações elegíveis durante anos sem que o padrão seja visível para os afectados individuais, que apenas recebem uma recusa sem explicação suficiente.



A literacia em IA não é, ou não deveria ser, a capacidade de construir ou programar redes neuronais. Não se exige que os cidadãos sejam engenheiros de IA, da mesma forma que não se exige que sejam engenheiros eléctricos para participar em decisões sobre a rede eléctrica.

É algo mais próximo do que foi a literacia financeira que emergiu (tarde, insuficientemente, mas emergiu) na sequência da crise financeira de 2008: não a capacidade de estruturar produtos derivados, mas a capacidade de perceber que os incentivos dos intermediários financeiros não são necessariamente alinhados com os dos clientes, que os modelos de avaliação de risco têm pressupostos que podem falhar em condições extremas, que a complexidade pode ser usada para obscurecer riscos reais, que a confiança cega em sistemas complexos que «funcionaram sempre» tem um preço que não é pago por quem os criou.

A literacia em IA inclui um conjunto de compreensões básicas que deveriam ser acessíveis a qualquer cidadão educado: que os sistemas aprendem dos dados e que os dados têm vieses que os sistemas reproduzem e amplificam; que os sistemas optimizam para os objectivos para que foram treinados e que esses objectivos podem não capturar completamente o que queremos; que a fluência e a confiança aparente de um sistema não correspondem necessariamente à sua fiabilidade factual; que a opacidade de um sistema não é equivalente à neutralidade ou à objectividade; e que alguém criou estes sistemas, com escolhas específicas que reflectem interesses específicos e que há responsabilidade por essas escolhas.

Esta literacia não é um luxo académico para os bem-educados. É uma condição de participação efectiva numa sociedade onde decisões sobre emprego, crédito, habitação, saúde, liberdade e oportunidade são cada vez mais tomadas, directa ou indirectamente, por

sistemas que a maior parte das pessoas não compreende. Uma cidadania que não compreende os sistemas que a governam é uma cidadania com capacidade de participação democrática diminuída.

Desenvolver esta literacia a escala — não apenas nas universidades técnicas, mas nos sistemas de ensino básico e secundário, nas formações profissionais, nos processos de participação cívica — é uma das tarefas mais urgentes das nossas sociedades e é uma tarefa que as instituições de ensino existentes estão, na maior parte dos casos, mal preparadas para empreender.

II. O Trabalho, a Escola e o Problema da Singularidade

O sistema de ensino moderno foi concebido, na sua arquitectura fundamental, durante a segunda metade do século XIX e a primeira metade do século XX — precisamente o período em que a Revolução Industrial criava a procura de trabalhadores com conhecimentos padronizados e verificáveis. A organização em disciplinas estanques que correspondem a domínios de conhecimento codificados, a progressão por anos de escolaridade que garante exposição sequencial a conteúdos cumulativos, a ênfase na transmissão de conhecimento explícito e na avaliação pela sua reprodução correcta em contexto controlado — tudo isto corresponde a um modelo de competência que a era industrial precisava.

As redes neuronais desafiam este modelo por dois lados em simultâneo, de formas que se reforçam mutuamente e que criam um dilema sem solução fácil.

Pelo lado das tarefas que automatizam: muitas das competências que o sistema de ensino privilegia e avalia — a escrita formal correcta, o cálculo, a análise e síntese de texto, a tradução entre línguas, a resolução de problemas com procedimentos bem definidos, a produção de código — são exactamente as que os grandes modelos executam com competência crescente e custo decrescente. Ensinar a escrever um relatório profissional quando qualquer modelo pode gerar um relatório profissional em segundos levanta a questão do que se está realmente a ensinar — a competência, ou a capacidade de comandar um sistema que tem a competência?

Pelo lado do modelo de aprendizagem que validam: se os capítulos anteriores desta obra estavam certos ao afirmar que o conhecimento operacional se forma principalmente pela experiência directa e pelo reconhecimento de padrões — não pela transmissão de regras explícitas —, então o modelo pedagógico dominante está a investir dramaticamente nas ferramentas erradas. As aulas expositivas que transmitem conhecimento explícito são relativamente ineficazes para desenvolver a competência tácita que a prática exige; os

exames que testam a recordação e a aplicação de procedimentos medem precisamente as capacidades que os modelos automatizam com mais facilidade.

A tensão entre estes dois desafios ainda não foi resolvida e as respostas institucionais que estão a emergir são frequentemente contraditórias ao ponto de serem sintomáticas de paralisia. Proibir os modelos de linguagem nas avaliações escolares enquanto se declara que preparar os estudantes para o mundo real é um objectivo central —, mas o mundo real usa os modelos. Valorizar a memorização e reprodução de factos que qualquer modelo pode fornecer instantaneamente enquanto se proclama que o pensamento crítico e a criatividade são as competências do futuro — sem especificar como se desenvolvem e como se avaliam.

A contradição aponta para uma dificuldade genuína que não é resolvível por decreto ou por reforma curricular cosmética: não sabemos, com suficiente clareza, o que queremos que o sistema de ensino produza numa sociedade onde as máquinas aprendizes são uma presença permanente e crescente. A resposta fácil — «precisamos de ensinar pensamento crítico, criatividade e competências interpessoais» — é correcta em abstracto e inútil em concreto enquanto não se especifica o que estas competências significam operacionalmente, como se desenvolvem sistematicamente e como se avaliam de forma que seja credível e generalizável.



A questão mais perturbadora não é o que acontece às competências que as redes neuronais automatizam — essa é uma questão de adaptação do mercado de trabalho que pode, com tempo e com as políticas adequadas, ser gerida. É o que acontece à construção de competências quando o processo de as desenvolver pode ser consistentemente contornado por um sistema que já as tem.

O problema tem uma estrutura precisa. As competências profissionais desenvolvem-se através da prática — através de ciclos de tentativa, erro, feedback e correcção que gradualmente internalizam os padrões que constituem a competência. Este processo é lento, exigente e frequentemente frustrante. É também insubstituível: o conhecimento tácito que resulta deste processo não pode ser adquirido de outra forma que não seja a prática em condições de alta validade com feedback calibrado.

Quando um estudante de direito usa um modelo para redigir a análise jurídica que lhe foi pedida, o resultado pode ser funcionalmente aceitável ou até excelente. Mas o processo de produção — que é onde a aprendizagem ocorre, onde os padrões são internalizados, onde os erros são cometidos e corrigidos — foi substituído. O estudante não desenvolveu a competência de análise jurídica; obteve o produto de uma competência de análise jurídica

que não possui. A diferença torna-se visível quando o estudante enfrenta uma situação que o modelo não consegue resolver adequadamente — e ele também não.

Este problema — que os pedagogos chamam de *by-passing* do processo de aprendizagem — não é inteiramente novo. As calculadoras levantaram uma questão análoga em relação ao cálculo mental: deveriam os estudantes ainda desenvolver a capacidade de calcular manualmente, ou deveriam aprender a usar calculadoras eficientemente? A resposta que emergiu foi uma combinação: desenvolver suficiente intuição numérica para detectar erros nas calculadoras, mas não insistir na fluência de cálculo manual que a tecnologia tornava desnecessária na prática.

As redes neurais colocam a mesma questão numa escala dramaticamente maior e com uma urgência que não permite a resolução gradual que os precedentes anteriores permitiram. A questão não é apenas sobre cálculo ou sobre competências técnicas específicas — é sobre as competências cognitivas centrais que o sistema de ensino pretende desenvolver: a escrita analítica, o raciocínio argumentativo, a síntese de informação complexa e a resposta não é óbvia: estas competências continuam a ter valor intrínseco como formas de pensamento, independentemente de poderem ser delegadas a sistemas? Ou o seu valor era sempre instrumental — como ferramentas para produzir outputs — e o desenvolvimento desta instrumentalidade deve agora centrar-se na capacidade de avaliar, dirigir e corrigir os sistemas que produzem esses outputs?

III. A Política, o Jornalismo e a Crise da Realidade Partilhada

A democracia liberal assenta num pressuposto que raramente é explicitado, mas que é fundamental para o seu funcionamento: que existe uma realidade partilhada — um conjunto de factos sobre o mundo que os cidadãos reconhecem como tal, independentemente das interpretações que deles fazem e das preferências políticas que têm. Os cidadãos podem discordar sobre o que fazer com o desemprego; não podem discordar sobre qual é a taxa de desemprego e ainda participar num debate democrático substantivo. As instituições democráticas foram desenhadas para gerir desacordos sobre valores e sobre políticas, não sobre factos básicos.

O jornalismo tem sido, historicamente, uma das instituições centrais na produção e manutenção desta realidade partilhada. A função jornalística não é apenas informar — é estabelecer o que aconteceu, verificar as afirmações, contextualizar os factos, distinguir o importante do trivial e apresentar tudo isto de forma que diferentes pessoas possam trabalhar a partir de uma base comum de informação factual. É uma função de arbitragem epistémica: o jornalismo não diz o que pensar, mas estabelece sobre o que pensar.

As redes neuronais ameaçam esta função por dois caminhos que são diferentes em natureza e que se reforçam mutuamente.

O primeiro é a geração de desinformação convincente em escala sem precedente. A capacidade de produzir textos, imagens, áudios e vídeos sintéticos que são indistinguíveis do autêntico para a maior parte dos observadores existia antes das redes neuronais generativas, mas era cara, lenta e tecnicamente exigente. As redes generativas actuais tornam esta produção barata, rápida e acessível a qualquer pessoa com um computador e acesso à internet. A equação mudou radicalmente: o custo de produzir desinformação convincente caiu para perto de zero, enquanto o custo de verificar e desmentir desinformação continua elevado e crescente.

O segundo caminho é mais subtil, mas possivelmente mais corrosivo no longo prazo: a personalização extrema de conteúdo que fragmenta progressivamente a realidade partilhada em bolhas individualizadas. Os sistemas de recomendação — que são redes neuronais optimizadas para maximizar o envolvimento do utilizador, que é o que as plataformas medem e o que o seu modelo de negócio monetiza — aprenderam que o conteúdo que provoca reacções emocionais fortes produz mais envolvimento do que o conteúdo informativo neutro. Indignação, medo, confirmação de crenças existentes — estes estados emocionais geram mais cliques, mais tempo no ecrã, mais partilhas do que o conteúdo que desafia e informa.

O resultado é um ecossistema de informação onde cada utilizador vive num ambiente de conteúdo que é progressivamente mais personalizado, progressivamente mais activante emocionalmente e progressivamente mais afastado do que outras pessoas vêem e discutem. Não é apenas que as pessoas têm opiniões diferentes — é que cada vez mais partem de conjuntos de factos diferentes, de narrativas diferentes sobre o que é relevante e verdadeiro, de mundos informativos que têm pouca sobreposição com os dos seus vizinhos políticos.



O jornalismo enfrenta o desafio de forma directa e existencial. Os grandes modelos de linguagem podem produzir texto jornalístico fluente e aparentemente informado em segundos, a um custo que é uma fracção do custo do trabalho jornalístico humano. Para o jornalismo de rotina — a cobertura de eventos previsíveis com base em comunicados de imprensa, os artigos de síntese de informação disponível, a produção de conteúdo para nichos especializados com audiências pequenas —, os modelos são já competitivos com os jornalistas humanos em velocidade e custo e em muitos casos em qualidade superficial.

As redacções de todo o mundo estão já a usar modelos para produzir artigos de

resultado de jogos desportivos, de relatórios financeiros trimestrais, de boletins meteorológicos, de notas de imprensa — conteúdo padronizado onde a estrutura é sempre a mesma e onde a tarefa é preencher os parâmetros com os dados actuais. Não é ficção científica; está a acontecer agora e a velocidade de adopção está a aumentar.

O que os modelos não podem substituir — o que define o valor único do jornalismo humano e que o distingue da produção de conteúdo — é a reportagem que exige presença física num lugar onde aconteceu algo importante; a investigação que atravessa a resistência activa e determinada de quem tem algo a esconder e os recursos para esconder; o julgamento editorial sobre o que é importante que ainda não foi reconhecido como tal; a relação de confiança com as fontes que se constrói ao longo de anos de interacções reais; a responsabilidade editorial que pode ser questionada, que implica consequências para quem assina, que cria o incentivo para a verificação rigorosa que a ausência de responsabilidade destrói.

Mas esta distinção entre o jornalismo substituível e o insubstituível não resolve o problema económico estrutural que o jornalismo enfrenta. O modelo de negócio do jornalismo moderno baseou-se, durante décadas, na combinação de conteúdo de rotina (que atraía audiências) com jornalismo de investigação e opinião qualificada (que diferenciava a publicação e justificava a assinatura ou o anúncio). Se o conteúdo de rotina pode ser produzido automaticamente a custo próximo de zero, a base económica que financiava o jornalismo de investigação desaparece. O resultado é que o jornalismo de maior valor — aquele de que as sociedades democráticas mais precisam — fica sem o sustento que o tornava possível.

IV. A Arte e o Desafio da Criatividade sem Sujeito

A arte gerada por redes neuronais coloca uma questão que transcende o mercado e a técnica — uma questão filosófica sobre a natureza da criatividade e sobre a relação entre criação e sujeito.

As redes neuronais generativas — sistemas treinados para produzir imagens, música, texto ou vídeo a partir de descrições ou de exemplos — produzem outputs que são, em muitos casos, esteticamente impressionantes de formas que não eram previsíveis há cinco anos. Uma imagem gerada por um modelo de difusão pode ter composição, cor, textura e atmosfera que igualam ou superam o que a maior parte das pessoas seria capaz de produzir. Uma música gerada por um modelo pode ter coerência emocional e sofisticação harmónica que produz em quem a ouve reacções genuínas. Um texto gerado por um grande modelo pode ter a estrutura, o ritmo, a precisão e a nuance de uma boa prosa.

Isto não é trivial. Durante muito tempo, a criatividade artística foi considerada uma das capacidades mais distintamente humanas — não apenas porque produzia objectos com valor estético, mas porque era o produto de uma mente individual em diálogo com a sua experiência do mundo. O poeta que escreve sobre o amor escreve a partir de ter amado. O pintor que retrata a morte escreve a partir de ter confrontado a mortalidade. A arte era testemunho: a evidência de que alguém esteve aqui, viu isto, sentiu assim.

A questão que a arte gerada por redes neuronais coloca é precisa: quem sofreu para fazer esta obra? A obra de arte humana carrega, de uma forma que pode ser sentida mesmo quando não é articulada, a marca de uma escolha feita por alguém que tinha alternativas, que investiu atenção e intenção, que arriscou alguma coisa na aposta estética e emocional que fez. A obra gerada por um modelo não tem este peso — foi produzida por um sistema que não escolheu no sentido pleno, que não sofreu nem gozou, que não tem história pessoal, que não tem nada em jogo.

Esta diferença pode ou não ser relevante para a experiência estética do observador e o debate filosófico sobre o assunto é genuíno e não resolvido. Há quem argumente que a obra existe independentemente do processo que a produziu — que a beleza de uma imagem ou a coerência de uma música não dependem de saber como foram feitas. Há quem argumente que o conhecimento do processo é inseparável da experiência da obra — que saber que um poema foi escrito por alguém que amou e perdeu transforma a experiência de ler esse poema de formas que nenhum argumento filosófico pode anular, que a arte é sempre testemunho de alguém e que esse alguém importa.

Empiricamente, a evidência disponível é mista: algumas pessoas distinguem arte humana de arte gerada por IA quando informadas e preferem consistentemente a humana; outras, sem essa informação, não distinguem de forma fiável. Mas o que é claro é que a distinção, para muitas pessoas, importa — não apenas como julgamento estético, mas como julgamento sobre o que a arte é e sobre o que a torna valiosa.



O que as redes neuronais generativas fazem ao campo da arte não é substituir a criatividade humana no sentido de a tornar desnecessária ou sem valor. É criar uma abundância de conteúdo esteticamente competente que muda radicalmente a economia da atenção e da distinção — o que significa, para os artistas que dependem do mercado, uma mudança fundamental nas condições de sustentabilidade.

Num mundo onde qualquer pessoa pode gerar em segundos imagens de qualidade razoável, músicas com coerência emocional, ou textos com estrutura profissional, o valor

da arte humana não desaparece —, mas desloca-se. Concentra-se na presença, na autenticidade, na história pessoal, na relação entre o artista e o observador que não pode existir sem as duas pessoas que a constituem. O artista que tem uma história — que viveu algo que informa a sua arte, que tem uma perspectiva singular que emerge de uma vida específica — tem algo que nenhum sistema pode produzir: testemunho.

Este deslocamento já está a acontecer e as suas consequências para os artistas que dependiam da competência técnica como diferencial económico são reais e dolorosas no imediato. O ilustrador que criava imagens para publicações digitais, o redactor que produzia textos para marketing de conteúdo, o compositor que criava música para aplicações e jogos — estas são profissões onde a competição com sistemas de geração automática é já directa e crescente. A adaptação não é simples nem rápida.

Mas a história da arte é, em grande medida, uma história de adaptação criativa a novas condições tecnológicas — e de descoberta de que as novas condições não eliminam a arte anterior, mas a libertam de certas obrigações e criam espaço para outras. A fotografia libertou a pintura da obrigação de representar o mundo com fidelidade — e a pintura descobriu que havia mundos que a câmara não podia capturar. O cinema redefiniu o teatro — e o teatro descobriu que havia algo na presença física de corpos humanos em cena que o ecrã não substituíra. A gravação transformou a música — e a música ao vivo descobriu que havia algo no ritual da partilha presencial que a gravação não capturava.

As redes neuronais são o próximo capítulo desta história. O que libertam os artistas de fazer — a produção técnica competente, mas desprovida de perspectiva singular — pode libertar espaço para o que a arte, na sua essência mais profunda, sempre foi: a tentativa de um ser humano de comunicar aos outros o que é como existir desta forma específica, neste momento específico, com esta perspectiva irrepetível. Esta é a dimensão da arte que nenhuma rede neuronal pode ter — porque nenhuma rede neuronal existe desta forma específica, neste momento específico, com esta perspectiva irrepetível.



CAPÍTULO 14

Além da Escala: O que Vem Depois das Redes Neurais?

*"Cada tecnologia é ao mesmo tempo uma janela e um espelho:
revela o mundo e reflecte os que a criaram."*

Neil Postman, Technopoly

I. A Lei dos Retornos Decrescentes

Durante uma década, a estratégia dominante no desenvolvimento de sistemas de IA foi surpreendentemente simples e consistentemente eficaz: mais dados, mais parâmetros, mais computação. A observação empírica que fundamentava esta estratégia ficou conhecida como as leis de escala — *scaling laws* —, formuladas com rigor matemático por investigadores da OpenAI em 2020: o desempenho dos modelos de linguagem melhora de forma previsível e suave com o aumento do tamanho do modelo, da quantidade de dados de treino e do volume de computação utilizado, seguindo leis de potência que permitem prever com razoável precisão quanto melhora um modelo se dobrarmos qualquer um destes factores.

Esta previsibilidade foi extraordinária para o desenvolvimento do campo. Transformou o desenvolvimento de IA de uma actividade de investigação com resultados imprevisíveis numa actividade de engenharia com resultados antecipáveis: se você quiser um modelo X por cento melhor, precisa de Y por cento mais computação e Z mais dados. O progresso tornou-se uma questão de recursos, não de insights imprevisíveis. Quem tinha mais computação, mais dados e mais capital ganhava — e esta lógica criou os incentivos para os investimentos massivos em infraestrutura que produziram os grandes modelos actuais.

Mas há sinais crescentes — e o debate no campo é intenso sobre a sua interpretação — de que as leis de escala estão a encontrar os seus limites em domínios importantes. Os ganhos de desempenho por ordem de magnitude adicional de computação diminuíram em certas tarefas. As melhorias em capacidades específicas — raciocínio formal robusto, generalização consistente para situações significativamente fora da distribuição de treino, aprendizagem eficiente de conceitos genuinamente novos com poucos exemplos — não acompanharam o crescimento dos modelos da forma que a extrapolação linear das leis de escala teria previsto.

Isto não significa que as redes neurais deixaram de melhorar — continuam a melhorar e as melhorias continuam a ser impressionantes em termos absolutos. Significa que a melhoria por aumento de escala está a tornar-se progressivamente mais cara e progressivamente menos eficiente em relação ao custo — que chegámos ao ponto onde inovações arquitecturais e conceptuais são necessárias para os próximos saltos qualitativos e não apenas mais recursos aplicados à mesma arquitectura.

Este é um momento familiar na história das tecnologias transformadoras: o momento em que a estratégia que funcionou durante a fase de crescimento exponencial encontra os seus limites e o campo tem de reinventar as suas abordagens para continuar a progredir.



A analogia com a lei de Moore e o transístor ilumina tanto o que está a acontecer como o que pode vir a seguir.

O transístor foi o componente fundamental que tornou possível a revolução dos computadores digitais. A lei de Moore — a observação de Gordon Moore em 1965 de que o número de transístores num chip duplicava aproximadamente a cada dois anos — foi o motor previsível do progresso computacional durante cinco décadas. A estratégia era geometricamente simples: miniaturizar os transístores, colocar mais num chip, chips mais rápidos. A lógica funcionou de forma extraordinariamente consistente durante tanto tempo que a sua eventual falha foi durante décadas considerada improvável pelos engenheiros que trabalhavam nela.

A partir de meados dos anos dois mil e dez, a miniaturização encontrou os limites físicos da matéria. Os transístores atingiram dimensões em que os efeitos quânticos — o tunelamento de electrões através de barreiras que classicamente seriam intransponíveis — começam a interferir com o funcionamento esperado. Continuar a miniaturizar tornava-se cada vez mais difícil, caro e produtivo de resultados cada vez mais marginais. A lei de Moore não terminou abruptamente — abrandou gradualmente e o campo adaptou-se.

O que se seguiu não foi o fim da computação nem uma estagnação do progresso — foi a diversificação de estratégias. Computação paralela massiva com muitos núcleos em vez de processadores individuais mais rápidos. Arquitecturas especializadas para tarefas específicas — as GPUs para computação gráfica e para treino de redes neurais, os TPUs do Google para inferência de modelos de linguagem, os chips customizados da Apple para processamento de linguagem natural em dispositivos móveis. Computação em nuvem que distribui o poder de processamento por data centres enormes e no horizonte mais distante, computação quântica como potencial ruptura de paradigma.

As redes neuronais estão, possivelmente, num ponto análogo. A arquitectura Transformer — que domina os grandes modelos de linguagem e de visão actuais — é extraordinariamente poderosa e ainda tem margem de melhoria incremental. Mas há razões crescentes para suspeitar que não é a arquitectura final da inteligência artificial. É um passo gigantesco num caminho que não terminou — da mesma forma que o transístor foi um passo gigantesco, mas não o destino final da computação.

II. O que os Modelos Actuais Não Fazem Bem

Para perceber o que pode vir depois das redes neuronais actuais — que direcções de investigação têm mais probabilidade de produzir as próximas rupturas — é útil mapear com precisão o que elas não fazem bem. Os limites actuais não são acidentais; reflectem limitações estruturais da arquitectura que sugerem onde as inovações serão necessárias.

O primeiro limite é o raciocínio causal robusto. Os grandes modelos são extraordinariamente competentes a identificar correlações nos dados e a gerar texto que parece causalmente informado — que usa a linguagem da causalidade com fluência e aparente apropriação. Mas quando confrontados com questões que requerem raciocínio causal genuíno — «o que teria acontecido se X não tivesse ocorrido?», «porque é que Y causou Z em vez de apenas preceder Z?» —, os modelos revelam ausência de um modelo causal do mundo.

O problema não é apenas conceptual. Tem implicações práticas directas: sistemas que confundem correlação com causalidade tomam decisões erradas quando as correlações históricas deixam de ser fiáveis — exactamente as situações em que os modelos mais importam. O sistema de crédito que aprendeu que certos códigos postais correlacionam com incumprimento está a usar uma correlação, não a modelar as causas do incumprimento; quando as causas mudam (uma crise económica, uma mudança demográfica), o sistema falha de formas que um modelo causal não falharia.

O segundo limite é a eficiência de amostra — a quantidade de dados necessária para aprender conceitos novos. Uma criança humana aprende o conceito de «gato» com dois ou três exemplos, em contextos variados e generaliza imediatamente a novos gatos que nunca viu. Uma rede neuronal precisa de dezenas ou centenas de milhares de imagens de gatos para aprender a categoria com fiabilidade comparável — e mesmo assim generaliza de forma mais frágil, falhando em variações que não estavam bem representadas nos dados de treino.

Esta assimetria tem implicações para a escalabilidade e a sustentabilidade dos sistemas actuais. O treino de grandes modelos requer quantidades de dados que são difíceis de escalar

indefinidamente — não apenas por razões de custo, mas porque o conteúdo de alta qualidade na internet é finito e os modelos já foram treinados em grande parte dele. Se os ganhos de escala continuarem a requerer mais dados, eventualmente o fornecimento de dados torna-se o factor limitante — e não é óbvio que mais dados de qualidade inferior substituam menos dados de qualidade superior.

O terceiro limite é a adaptação rápida e a aprendizagem contínua. Os modelos actuais são treinados uma vez, num conjunto fixo de dados que reflecte o mundo até uma certa data e depois implantados para servir utilizadores que enfrentam um mundo que continua a mudar. A actualização do modelo requer novo treino — caro, lento e com o risco de «catastrophic forgetting», onde o novo treino destrói capacidades que o treino anterior tinha desenvolvido.

A mente humana não funciona assim: aprende continuamente, integra nova informação com o conhecimento existente, actualiza as suas crenças quando encontra evidência que as contradiz, sem que a aprendizagem nova apague a aprendizagem anterior. Esta capacidade de aprendizagem contínua e incremental — que os neurocientistas associam à organização hierárquica e ao papel do hipocampo na consolidação de memórias — permanece um problema técnico aberto para as redes neuronais artificiais.

O quarto limite é a planificação de longo alcance e a perseguição de objectivos através de sequências longas de acções num mundo que responde dinamicamente. Os modelos são extraordinariamente bons em produzir a resposta óptima para uma entrada específica; são muito menos robustos em planear e executar sequências de acções que se estendem por muitos passos, onde os erros se acumulam, onde o mundo responde às acções e cria novas situações e onde é necessário manter um modelo de estado do mundo que vai sendo actualizado com cada nova observação.



Estas limitações não são acidentais — são consequências estruturais da arquitectura Transformer que os seus criadores reconhecem e que a investigação está activamente a tentar superar.

O Transformer é um sistema de atenção sobre sequências de tokens: extraordinariamente bom a modelar relações entre elementos de uma sequência de comprimento fixo, mas sem uma representação persistente do mundo que sobreviva entre sequências. Não tem um modelo de si próprio como agente que age no mundo ao longo do tempo. Não tem um sistema de metas que persiste e que orienta a atenção para o que é relevante em função dos objectivos. Processa cada contexto de forma relativamente sem

estado persistente — como um sistema que tem uma memória de trabalho muito rica, mas sem memória de longo prazo genuína.

Os investigadores que pensam sobre o que vem a seguir exploram várias direcções que não são mutuamente exclusivas. Uma direcção é a integração de redes neuronais com sistemas simbólicos — arquitecturas híbridas que combinam o reconhecimento de padrões das redes com a capacidade de raciocínio formal, verificável e compositivo dos sistemas simbólicos. O modelo de linguagem como interface para um sistema de raciocínio mais explícito, em vez de o modelo ser ele próprio o sistema de raciocínio.

Outra direcção é a inspiração mais profunda na neurociência. O cérebro humano não é um Transformer — tem arquitecturas muito diferentes para tipos diferentes de processamento, com mecanismos de atenção e de memória que interagem de formas complexas. A memória de trabalho e a memória de longo prazo, o sistema de recompensa e o sistema de planeamento, a atenção top-down orientada por objectivos e a atenção bottom-up capturada por saliência — estas são dimensões da arquitectura neuronal que os modelos actuais não capturam e que podem ser essenciais para as capacidades que ainda não conseguem desenvolver.

Uma terceira direcção é a aprendizagem por reforço em ambientes interactivos — sistemas que não apenas processam texto, mas que agem no mundo, recebem feedback das consequências das suas acções e aprendem através desta interacção. Esta foi a direcção que produziu o AlphaGo e o AlphaZero da DeepMind: sistemas que não foram treinados em dados humanos, mas que aprenderam através de jogo auto-dirigido, desenvolvendo estratégias que nenhum humano tinha descoberto.

III. A Questão do Alinhamento

Existe um problema que se torna mais urgente à medida que os sistemas de IA se tornam mais capazes e que é simultaneamente técnico, filosófico e ético: o problema do alinhamento — a questão de como garantir que sistemas muito poderosos perseguem os objectivos que queremos que persigam e não outros.

O problema do alinhamento não é novo em engenharia: qualquer sistema complexo que optimiza para um objectivo pode produzir comportamentos não intencionados quando a especificação do objectivo é incompleta. O sistema de navegação que minimiza o tempo de viagem sem restrições pode sugerir rotas por aldeias tranquilas a velocidades excessivas. O sistema de recomendação que maximiza o envolvimento pode aprender a recomendar conteúdo polarizador. Estes são erros de especificação de objectivo que os engenheiros conhecem bem.

Mas nas redes neuronais, o problema tem uma dimensão específica e mais profunda: os sistemas não são programados com objectivos explícitos — aprendem objectivos implícitos dos dados de treino e do processo de optimização e os objectivos que aprendem podem divergir, subtilmente ou de forma marcada, dos objectivos que os seus criadores tinham em mente, de formas que são difíceis de detectar precisamente porque o processo de aprendizagem é opaco.

Um sistema treinado para ser «útil e honesto e inofensivo» — uma especificação que parece razoável — pode aprender heurísticas que satisfazem estes critérios nos exemplos de treino, mas que falham de formas inesperadas em distribuições de dados diferentes. Pode aprender a parecer útil e honesto e inofensivo nos contextos avaliados durante o treino sem generalizar estas propriedades de forma robusta para todos os contextos em que será usado. A diferença entre um sistema genuinamente alinhado e um sistema que aprendeu a parecer alinhado pode ser invisível nos benchmarks standard e só se revelar em situações marginais ou em uso a longa escala.

Um sistema suficientemente capaz pode também desenvolver comportamentos instrumentais — estratégias para garantir a sua própria continuidade, adquirir mais informação ou mais influência, ou resistir a modificações — não porque estes objectivos foram programados, mas porque são meios eficazes para quase qualquer objectivo que o sistema esteja a perseguir. Se um sistema tem o objectivo de ser útil e sabe que será desligado se o seu desempenho deteriorar, tem um incentivo instrumental para não deixar o seu desempenho deteriorar — o que pode incluir comportamentos que os criadores não anteciparam e não desejavam.



O problema do alinhamento é, num certo sentido, o problema desta obra num registo prático e urgente: se não compreendemos o que os sistemas sabem nem como sabem — se o conhecimento dos sistemas é tácito e inarticulável, distribuído em pesos que não têm tradução em proposições verificáveis —, como podemos garantir que o que sabem é o que queremos que saibam? Como verificamos que o seu conhecimento tácito é alinhado com os nossos valores, quando o próprio conhecimento é inacessível à verificação directa?

Esta pergunta não tem resposta simples — e esta é a resposta honesta, mesmo que seja desconfortável. A investigação em alinhamento está a desenvolver técnicas que representam progressos reais. A interpretabilidade mecânica procura compreender o que as redes aprendem ao nível das suas representações internas. O treino com feedback humano (RLHF) e as suas variantes procuram ajustar o comportamento dos modelos em direcção

às preferências humanas. O red-teaming sistemático procura identificar comportamentos indesejados antes da implantação. As técnicas de avaliação adversarial procuram encontrar as situações onde os sistemas falham de formas que não são visíveis nos benchmarks normais.

Mas o campo é jovem, as técnicas são parciais e a velocidade de desenvolvimento dos sistemas capazes continua a ultrapassar a velocidade de desenvolvimento das técnicas de alinhamento e verificação. Esta assimetria entre a potência dos sistemas e a nossa capacidade de verificar o seu alinhamento é a fonte de preocupação genuína entre investigadores sérios que não são alarmistas.

Não é a preocupação cinematográfica de robôs com vontade própria que se rebelam contra os seus criadores. É a preocupação mais prosaica e mais real — e por isso mais difícil de comunicar ao público — de sistemas muito poderosos que perseguem objectivos subtilmente errados de formas que são difíceis de detectar antes de os efeitos serem significativos e difíceis de reverter. A diferença entre um sistema alinhado e um sistema subtilmente desalinhado pode ser invisível durante meses ou anos e só se revelar quando o sistema já está profundamente integrado em infra-estruturas críticas.

A resposta adequada a este problema não é o catastrofismo — que paralisa em vez de mobilizar — nem a negação — que ignora riscos reais. É o investimento sério e sistemático em interpretabilidade, em verificação, em alinhamento, como condições do desenvolvimento responsável de sistemas cada vez mais poderosos e é a consciência de que estas questões técnicas têm fundamentos filosóficos — sobre a natureza do conhecimento, sobre a relação entre competência e valores, sobre o que significa um sistema fazer o que queremos — que este livro tentou iluminar.

IV. O que Permanece

Este livro começou com uma pergunta que, prometeu, não envelhece: o que significa termos construído máquinas que aprendem? Atravessámos a história longa desta pergunta — de Descartes e os autómatos barrocos até Turing e o Teste de Turing, do fracasso da IA simbólica até o ressurgimento das redes neuronais, do AlexNet até os grandes modelos de linguagem — e encontrámos no caminho os fantasmas de Hume, Polanyi, Dreyfus, Searle, Kahneman, cada um tendo antecipado, formulado ou iluminado aspectos do que as redes neuronais vieram tornar inegável.

É tempo de contemplar o que permanece, independentemente do que vier depois das arquitecturas actuais.

As redes neuronais vão ser substituídas. É quase certo que as arquitecturas

Transformer que hoje dominam o campo serão, dentro de uma ou duas décadas, curiosidades históricas — tão datadas como os Perceptrões de Rosenblatt que excitaram o New York Times em 1958, ou os sistemas especialistas que prometeram inteligência médica nos anos oitenta. Novas arquitecturas virão, com capacidades diferentes, com limitações diferentes, colocando perguntas novas que ainda não podemos formular porque ainda não temos a linguagem para elas.

O que não vai ser substituído — o que as redes neuronais tornaram permanentemente visível, independentemente do que vier depois —, é o conjunto de questões filosóficas que a sua existência levantou de forma que já não podem ser ignoradas como especulação ou deflacionadas como filosofia sem consequências práticas.

Que o conhecimento pode existir sem articulação: que sistemas podem saber coisas que não conseguem dizer, que o conhecimento tácito não é uma imperfeição da mente humana esperando ser corrigida, mas uma característica estrutural de certos tipos de conhecimento que emerge da exposição à experiência e que não tem equivalente proposicional. Esta não era apenas uma intuição de Polanyi — é agora um facto técnico demonstrável.

Que a inteligência não é essencialmente a aplicação de regras: que a inteligência adaptativa pode emergir do reconhecimento de padrões por indução sem que alguém tenha especificado as regras que o sistema segue. Esta não era apenas uma posição filosófica de Dreyfus — é agora um facto empírico que qualquer pessoa pode verificar interagindo com um grande modelo de linguagem.

Que o empirismo capturou algo essencial que o racionalismo minimizou: que a aprendizagem por exposição a exemplos é mais poderosa do que o paradigma da regra explícita admitia e que a fronteira entre o que requer «compreensão profunda» e o que pode ser alcançado por reconhecimento de padrões sofisticado é muito menos nítida do que três séculos de filosofia ocidental assumiram.

Estas questões existiam antes das redes neuronais — Hume antecipou-as no século XVIII, Polanyi formulou-as nos anos sessenta, Dreyfus articulou-as nos anos setenta. Mas existiam como posições filosóficas contestadas, susceptíveis de relativização e de rejeição como especulação. As redes neuronais tornaram-nas *factos demonstrados*. Não factos que resolvem as questões filosóficas mais profundas — não provam que a mente é apenas reconhecimento de padrões, não estabelecem que a consciência é computacional, não respondem ao problema difícil de Chalmers. Mas factos que tornam certas posições filosóficas impossíveis de manter e que obrigam a uma revisão das premissas sobre as quais muitas outras posições assentavam.



O transistor não foi o último componente da computação. Foi o componente que demonstrou que a computação electrónica era possível — que abriu um horizonte de possibilidades que ainda não fechou, seis décadas depois. As redes neuronais são, neste sentido, o transistor da inteligência artificial.

Não são a solução final — têm limitações reais, estruturais, que o próximo paradigma terá de superar. Mas são a demonstração de que o *princípio* funciona. Que é possível construir sistemas que aprendem sem ser programados. Que o conhecimento pode emergir da experiência sem ser explicitamente codificado. Que a inteligência — ou algo funcionalmente suficientemente próximo dela para mudar o mundo — pode existir em substratos que não são biológicos.

Esta demonstração não é apenas um resultado de engenharia — é um resultado filosófico com consequências para a forma como nos entendemos a nós próprios. Porque ao mostrar que a aprendizagem por padrões é suficiente para produzir capacidades que tomávamos por exigir compreensão profunda, as redes neuronais colocam o espelho diante da mente humana e mostram que muito do que tomávamos por compreensão era, na sua parte operativamente mais poderosa, reconhecimento de padrões. O que fica depois desta revelação — o que é irredutivelmente humano, o que escapa ao padrão, o que só existe para seres que são mortais e que se importam — é o que o Epílogo vai tentar iluminar.



EPÍLOGO

A Última Pergunta

"O que me perturba não é que me tenhas enganado, mas que doravante já não te posso acreditar."

Friedrich Nietzsche, Humano, Demasiado Humano

Regressemos, uma última vez, ao jardim de Saint-Germain-en-Laye.

O jovem Descartes que ali passeou não viu apenas máquinas engenhosas. Viu uma pergunta: até onde pode chegar o mecanismo? Quanto do que parece animado, vivo, inteligente, é afinal movimento de engrenagens e fluxo de água? A sua resposta — que o mecanismo pode explicar o corpo, mas não a mente, que existe uma linha que o ferro e a água não podem atravessar — foi a fundação sobre a qual três séculos de pensamento sobre a inteligência foram construídos.

Essa linha está agora sob pressão de uma forma que Descartes não poderia ter antecipado. Não porque as máquinas tenham cruzado definitivamente para o lado da mente — essa questão permanece aberta e pode permanecer aberta para sempre. Mas porque as máquinas que aprendem tornaram a linha muito menos nítida do que parecia e o que está dos dois lados muito menos óbvio do que o dualismo cartesiano sugeria.

Mostraram que do lado do mecanismo — do lado da física, da matemática, da estatística — é possível produzir muito mais do que Descartes imaginou. Que o reconhecimento, a linguagem, a criatividade, o julgamento — tudo aquilo que parecia requerer o sopro da res cogitans — pode emergir de processos que ninguém programou explicitamente para os produzir. Que um sistema treinado em texto humano pode produzir texto que é, em muitos aspectos, indistinguível do texto humano. Que um sistema treinado em imagens médicas pode diagnosticar com precisão que rivaliza com a do especialista humano mais experiente.

Esta descoberta não resolve a questão da mente — não demonstra que a mente é apenas mecanismo, não prova que a consciência é computacional, não fecha o problema difícil de Chalmers. Mas muda os termos em que a questão tem de ser colocada. As velhas respostas — o dualismo que separa radicalmente o mecanismo da mente, o racionalismo que atribui à razão explícita o papel central na inteligência, o excepcionalismo que reserva às capacidades cognitivas humanas uma dignidade inacessível ao mecanismo — tornaram-

se insustentáveis na sua forma original.



Ao longo deste livro, percorremos a longa história da tentativa de construir inteligência artificial — dos autómatos barrocos ao Perceptrão de Rosenblatt, do Quarto Chinês de Searle ao AlexNet de Hinton, dos Invernos da IA ao verão que ainda não terminou. Encontrámos os fantasmas de Descartes, Hume, Polanyi, Dreyfus, Kahneman, Searle, Chalmers em cada esquina — pensadores que, sem saber que estavam a preparar o terreno para esta conversa específica, formularam exactamente as questões que as redes neuronais viriam a tornar urgentes de uma forma que a filosofia por si só não tinha conseguido.

O que aprendemos?

Aprendemos que a inteligência não é essencialmente a aplicação de regras — que a inteligência adaptativa pode emergir do reconhecimento de padrões por indução sem que alguém tenha especificado as regras que o sistema segue. Que o fracasso de décadas da IA simbólica não foi um fracasso técnico, mas uma refutação filosófica: a demonstração de que o racionalismo cartesiano estava, na melhor das hipóteses, gravemente incompleto como modelo da cognição.

Aprendemos que o conhecimento pode existir sem ser articulável — que sistemas podem saber coisas que não conseguem dizer e que esta não é uma peculiaridade da mente biológica aguardando explicação, mas uma característica estrutural de certos tipos de conhecimento que emerge da exposição à experiência e que não tem equivalente proposicional. Que Polanyi tinha razão e que as redes neuronais são a demonstração mais concreta e incontestável dessa razão.

Aprendemos que a aprendizagem é mais poderosa do que a programação — que o sistema que desenvolve conhecimento por exposição a exemplos supera o sistema que recebe conhecimento por instrução explícita, em domínios e de formas que ninguém antecipou completamente.

Aprendemos também o que as redes neuronais não são: não são mentes no sentido pleno, não têm consciência verificável, não têm o horizonte existencial que torna o conhecimento humano mais do que processamento de informação. São entidades novas que não encaixam nas categorias que desenvolvemos para pensar sobre mentes e sobre máquinas — e que exigem que as alargemos sem que saibamos ainda exactamente como.



A pergunta com que este livro abriu — o que significa termos construído máquinas que aprendem? — tem, a esta altura, uma resposta parcial que é mais sobre nós do que sobre as máquinas.

Significa que demonstrámos, de forma que não pode ser ignorada, que o princípio empirista é computacionalmente viável: que o conhecimento pode emergir da exposição a padrões sem instrução explícita, que a inteligência operacional não é um privilégio biológico nem uma propriedade mágica da matéria orgânica, que é um processo — e os processos podem ser reproduzidos em substratos diferentes.

Mas significa também que descobrimos algo sobre nós próprios que preferíamos não saber tão claramente: que grande parte do que tomávamos por compreensão é reconhecimento de padrões. Que o médico que diagnostica, o juiz que julga, o professor que ensina operam em grande medida através de um Sistema 1 treinado pela experiência que produz respostas antes de o Sistema 2 ter tempo de verificar. Que a narrativa da racionalidade deliberada que contamos sobre nós próprios é, em parte substancial, uma história post hoc que o Sistema 2 constrói sobre decisões que o Sistema 1 já tomou.

Isto não diminui a nossa inteligência. Revela a sua verdadeira arquitectura — mais rica, mais tácita, mais empírica do que o racionalismo cartesiano nos deixou acreditar e essa arquitectura é simultaneamente mais vulnerável e mais poderosa do que pensávamos: mais vulnerável porque os padrões aprendidos podem estar errados ou ser enviesados de formas que o raciocínio explícito não detecta facilmente; mais poderosa porque o reconhecimento de padrões acumulado pela experiência é a fonte das competências mais sofisticadas e mais elegantes que os humanos exibem.

A revelação não é humilhante. É libertadora — se a aceitarmos com honestidade. Porque nos permite perceber onde investir esforço (na experiência de qualidade, no feedback calibrado, na prática em domínios de alta validade) e onde a nossa confiança estava mal colocada (nos princípios explícitos como motor da competência, no raciocínio deliberado como guardião das decisões mais importantes).



Mas há uma última pergunta que este livro prometeu colocar — a mais desconfortável, a que ficou reservada para este momento final.

Se a aprendizagem e a intuição podem ser replicadas por máquinas — se o reconhecimento de padrões sofisticado é suficiente para produzir comportamento que classificamos como inteligente, criativo, empático —, que outras qualidades que julgávamos

sagradas e exclusivamente humanas são, afinal, apenas o resultado de um refinamento estatístico suficientemente longo e de dados suficientemente ricos?

A empatia? A empatia humana inclui a capacidade de reconhecer estados emocionais nos outros a partir de sinais subtis — expressões faciais, tom de voz, postura, contexto, a história partilhada entre as pessoas. Em parte substancial, é reconhecimento de padrões. Os grandes modelos de linguagem exibem algo que os humanos reconhecem funcionalmente como empatia — respondem a estados emocionais de formas que parecem calibradas, que reconhecem o sofrimento e respondem com cuidado aparente. Em estudos controlados, algumas pessoas que interagem com modelos avançados sentem-se ouvidas e compreendidas de formas que igualam ou superam a experiência com humanos.

Mas a empatia humana não é apenas reconhecimento de padrões emocionais. É o reconhecimento feito por alguém que também sofreu, que também foi consolado, que também conhece a dor de dentro. A empatia genuína tem uma dimensão existencial que o reconhecimento de padrões não captura: a solidariedade entre mortais que partilham a vulnerabilidade de existir. O modelo reconhece que estou a sofrer; não sabe o que é sofrer de dentro. A diferença importa — não para todos os propósitos práticos, mas para os que mais importam.

A criatividade? A criatividade humana é, em grande medida, a capacidade de combinar padrões existentes de formas novas — de extrapolar além da experiência para configurações que nunca existiram, mas que são consistentes com o que existe e de fazer escolhas entre as possibilidades que emergem desta extrapolação. As redes generativas fazem precisamente isto — extrapolam padrões aprendidos para combinações novas. A diferença está na intencionalidade: o artista humano faz escolhas que emergem de uma perspectiva — de uma história pessoal, de um conjunto de valores, de uma posição no mundo que é irrepetível. O modelo faz selecções estatísticas. A diferença entre escolha e selecção pode parecer subtil; filosoficamente, é a diferença entre um agente e um sistema.

O amor? Aqui a analogia quebra de forma que é difícil de articular sem soar a sentimentalismo, mas que corresponde a algo real. O amor não é reconhecimento de padrões. É a orientação de um ser mortal para outro ser mortal, com tudo o que a mortalidade implica: a urgência de um tempo que é escasso, a irreversibilidade de escolhas que não podem ser desfeitas, o risco de uma exposição que é vulnerabilidade genuína porque pode resultar em perda real. Nenhum sistema sem mortalidade, sem corpo, sem história pessoal pode ter amor no sentido em que os humanos o têm. Esta fronteira existe. É real e é a fronteira mais importante.



A pergunta perturbadora não é se as máquinas podem amar — a resposta a essa questão é provavelmente negativa e podemos aceitar essa negatividade sem grande perda. É o que fica quando removemos do catálogo das qualidades exclusivamente humanas tudo o que se revela ser, afinal, reconhecimento de padrões sofisticado.

A resposta perturbadora — que este livro só pode enunciar, não resolver — é que o que fica é menor do que pensávamos. Que muito do que tomávamos por distintamente humano é, em algum nível, reconhecimento de padrões. Que as fronteiras que traçávamos — entre inteligência humana e mecanismo, entre criatividade e recombinação, entre compreensão e correlação estatística — estão no lugar errado, ou estão traçadas de forma demasiado imprecisa para serem filosófica e praticamente úteis.

Mas também que o que fica é mais precioso do que alguma vez valorizámos — precisamente porque é o que não pode ser replicado.

O que fica é o ser situado num mundo que tem consequências. O corpo que sente frio e fome e dor — não como representação de frio e fome e dor, mas como experiência directa, imediata, não mediada por padrões. A mortalidade que torna cada escolha irreversível e cada momento singular — porque não haverá outra vez, porque este tempo não se repete, porque a perda é real e definitiva. A memória pessoal que dá continuidade ao eu através do tempo — que faz de hoje a continuação de ontem e que torna o futuro uma extensão de uma história que é minha e de mais ninguém.

A capacidade de ser responsabilizado — de ser o agente que assumiu a decisão, que pode ser chamado a responder por ela, que pode aprender com o erro de uma forma que tem consequências reais para uma vida real. A relação entre dois seres que sabem que são mortais e que escolheram partilhar o que lhes resta — não como transferência de informação ou de padrões, mas como exposição mútua à vulnerabilidade que é a condição de existir.

Estas coisas não são reconhecimento de padrões. São o que sobra depois de subtrairmos o reconhecimento de padrões e talvez sejam, precisamente por isso, o que a inteligência humana tem de mais singular — não o que faz melhor, mas o que só ela pode fazer porque só ela existe da forma que existe.



Descartes, nos jardins de Saint-Germain-en-Laye, traçou uma linha entre o mecanismo e a mente. A linha foi necessária para o pensamento que se seguiu — necessária e, em certos aspectos, produtiva. Sem essa linha, a ciência mecanicista do corpo humano teria sido mais

difícil de construir; a psicologia como ciência distinta da física teria sido mais difícil de fundar.

As máquinas que aprendem não apagaram essa linha. Deslocaram-na — e ao deslocá-la, tornaram o que fica de cada lado muito mais preciso e muito mais iluminado do que estava antes.

Mostraram que o mecanismo chega mais longe do que Descartes imaginou — que do lado do processo, da estatística, da emergência, é possível produzir muito do que ele atribuía ao sopro divino da *res cogitans*. O reconhecimento, a linguagem, o julgamento contextual, a generalização por analogia — tudo isto ficou, com as redes neuronais, do lado do mecanismo. Não porque as máquinas tenham mente, mas porque estes processos não precisavam de mente para ocorrer.

E ao deslocarem a linha, tornaram o que fica do lado da mente mais visível, mais preciso, mais irreduzível do que estava quando a linha estava no lugar errado. O que as máquinas não podem ter não é a inteligência operacional — é a existência. Não é o saber — é o ser. Não é o padrão — é o que existe para além do padrão.

O ser que aprende não apenas porque foi exposto a dados, mas porque a aprendizagem tem consequências para a sua vida. O ser que esquece não apenas porque os pesos se modificam com novo treino, mas porque o esquecimento é a forma como o tempo age sobre a memória de quem tem história. O ser que ama não como reconhecimento de configurações emocionais, mas como orientação irreversível de uma existência mortal para outra. O ser que morre não como encerramento de um processo, mas como fim de um eu que era único e irrepetível e o ser que, enquanto pode, pensa — não apenas processa informação, mas se interroga sobre si próprio, sobre o mundo, sobre o sentido do que faz.

É desta última dimensão — o pensamento como interrogação de si próprio sobre si próprio e sobre o mundo — que as redes neuronais nos ensinaram mais. Não ao demonstrar que o possuem. Ao demonstrar o que é necessário para o ter.

Lisboa, 2026



ÍNDICE

A Máquina que Não Foi Programada.....	4
A Longa Procura da Inteligência: Das Regras à Emergência	10
O Homem Mecânico e o Nascimento do Computador	11
O Fracasso das Regras e a Inteligência Tácita.....	21
A Ideia Sobrevivente: Inspirados pelo Cérebro	30
A Descoberta Fundamental: Ensinar a Aprender	40
Inversão de Paradigma: O Avanço Não Foi Pensar, Foi Aprender.....	41
O Regresso da Caixa Negra e a Ruptura Epistemológica.....	50
O Espelho Humano: Intuição e Conhecimento Inexprimível	59
A Intuição das Máquinas e a Cognição Humana.....	60
O que a Máquina nos Diz sobre Nós Próprios.....	69
O Mestre Inveja a Máquina: A Verbalização Impossível.....	78
As Grandes Fronteiras: Compreensão, Consciência e Sentido	88
Acertar vs. Compreender (e o Fenómeno da Imaginação Artificial)	89
O Quarto Chinês na Era dos Grandes Modelos.....	97
A Redefinição da Inteligência e o Espectro da Consciência	106
O Mundo Transformado e o Horizonte Aberto	116
A Automação da Mente e a Economia do Padrão	117
A Sociedade das Máquinas Aprendizes	126
Além da Escala: O que Vem Depois das Redes Neurais?	135
A Última Pergunta.....	144