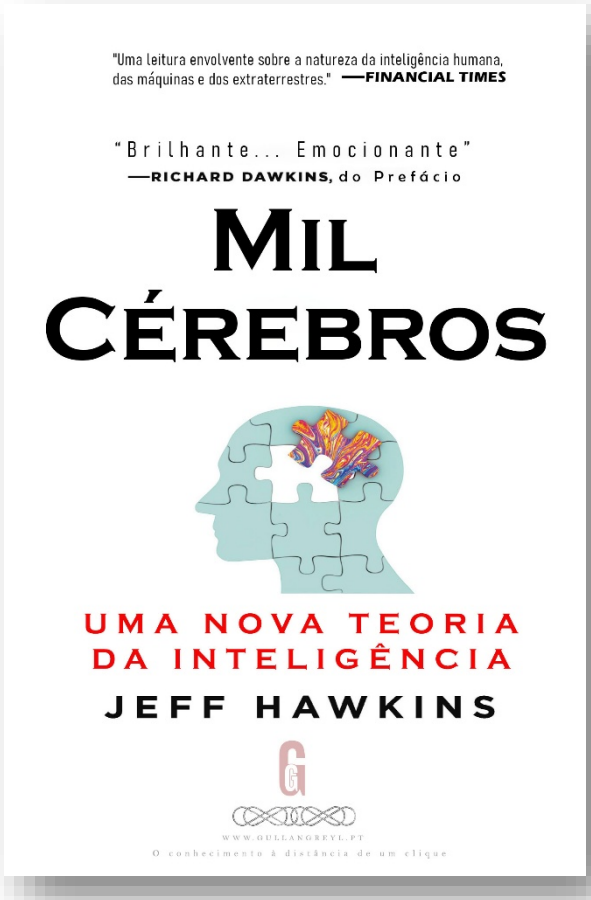


MIL CÉREBROS

Uma Nova Teoria da Inteligência

Digitalizado, Corrigido e Adaptado por Gullan Greyl
http://www.gullangreyl.pt

1ª Edição, 2021

14-07-2025

Síntese

Apesar de todos os avanços da neurociência, fizemos poucos progressos na sua maior questão: como é que as células simples no cérebro criam inteligência? Jeff Hawkins e a sua equipa descobriram que o cérebro utiliza estruturas semelhantes a mapas para construir um modelo do mundo – não apenas um modelo, mas centenas de milhares de modelos de tudo o que conhecemos. Esta descoberta permite a Hawkins responder a questões importantes sobre a forma como percebemos o mundo, por que temos um sentido de identidade e a origem do pensamento de alto nível. Mil Cérebros anuncia uma revolução na compreensão da inteligência. É um livro que reflete grandes ideias, em todos os sentidos da palavra. Um dos Melhores Livros de 2021 do Financial Times. Um dos Cinco Livros Favoritos de Bill Gates em 2021.

Índice

Prefácio de 1

RICHARD DAWKINS 1

PARTE 1 7

UMA NOVA COMPREENSÃO DO CÉREBRO 7

CAPÍTULO 1 18

Cérebro Velho – Cérebro Novo 18

CAPÍTULO 2 29

A Grande Ideia de Vernon Mountcastle 29

CAPÍTULO 3 37

Um Modelo do Mundo na Sua Cabeça 37

Aprendizagem Através do Movimento 42

Dois Princípios da Neurociência 44

Princípio Número Um: Pensamentos, Ideias e Perceções São a Atividade dos Neurónios 45

Princípio Número Dois: Tudo o que Sabemos Está Armazenado nas Ligações Entre os Neurónios 46

CAPÍTULO 4 48

O Cérebro Revela os Seus Segredos 48

Descoberta Número Um: O Neocórtex Aprende um Modelo Preditivo do Mundo 49

Descoberta Número Dois: As Previsões Ocorrem no Interior dos Neurónios 50

Descoberta Número Três: O Segredo da Coluna Cortical São Referências 57

CAPÍTULO 5 65

Mapas no Cérebro 65

Um Conto Evolutivo 66

Mapas no Cérebro Antigo 67

Mapas no Novo Cérebro 69

Mapas Enormes em Espaços Minúsculos 71

Mapas Numa Coluna Cortical 73

Orientação 75

CAPÍTULO 6 77

Conceitos, Linguagem e Pensamento de Alto Nível 77

Todo o Conhecimento é Armazenado em Quadros de Referência 79

Vias do “Quê” e “Onde” 82

Quadros de Referência para Conceitos 84

Método dos Loci 86

Estudos com Humanos Usando fMRI 86

Matemática 89

Política 91
Língua 92
Especialização 96
CAPÍTULO 7 98
A Teoria dos Mil Cérebros da Inteligência 98
A Visão Existente do Neocórtex 98
A Nova Visão do Neocórtex 103
Onde Está Armazenado o Conhecimento no Cérebro? 104
A Solução para o Problema da Ligação 106
Como é que a Votação é Realizada no Cérebro? 109
Estabilidade da Perceção 110
Atenção 112
Hierarquia na Teoria dos Mil Cérebros 114
PARTE 2 120
INTELIGÊNCIA DE MÁQUINA 120
CAPÍTULO 8 125
Por Que É Que Não Existe “Eu” na IA 125
Duas Vias para a AGI 127
O Cérebro como Modelo para a IA 130
Da Inteligência Artificial Dedicada à Inteligência Artificial Universal 132
Quando é que Algo é Inteligente? 135
Exemplos de Quadros de Referência 139
CAPÍTULO 9 143
Quando as Máquinas São Conscientes 143
Consciência 144
Qualia 146
A Neurociência da Consciência 150
Consciência Artificial 151
O Mistério da Vida e o Mistério da Consciência 152
CAPÍTULO 10 154
O Futuro da Inteligência de Máquina 154
As Máquinas Inteligentes Não Serão Como os Humanos 155
Aprendizagem versus Clonagem 167
As Aplicações Futuras da Inteligência Artificial São Desconhecidas 168
CAPÍTULO 11 171
Os Riscos Existenciais da Inteligência Artificial 171

A Ameaça da Explosão de Inteligência ... 173
A Ameaça do Desalinhamento de Objetivos ... 176
O Contra-Argumento ... 178
PARTE 3 ... 182
INTELIGÊNCIA HUMANA ... 182
CAPÍTULO 12 ... 184
Crenças Falsas ... 184
Vivemos numa Simulação ... 185
Crenças Falsas ... 189
Modelos Virais do Mundo ... 191
Modelos do Mundo Falsos e Virais ... 192
Linguagem e a Propagação de Crenças Falsas ... 194
CAPÍTULO 13 ... 197
Os Riscos Existenciais da Inteligência Humana ... 197
Os Riscos do Cérebro Antigo ... 198
Crescimento Populacional e Alterações Climáticas ... 200
Como o Neocórtex Pode Frustrar o Cérebro Antigo ... 203
O Risco das Crenças Falsas ... 204
A Grande Ideia ... 208
CAPÍTULO 14 ... 211
Unindo Cérebros e Máquinas ... 211
Por Que Sentimos Estar Presos no Corpo ... 212
Fazer o Upload do Seu Cérebro ... 214
Fundir o Cérebro com um Computador ... 218
CAPÍTULO 15 ... 221
Planeamento Patrimonial para a Humanidade ... 221
Mensagem numa Garrafa ... 224
Deixa as Luzes Acesas ... 226
Wiki Terra ... 232
CAPÍTULO 16 ... 236
Genes versus Conhecimento ... 236
Tornar-se uma Espécie Multiplanetária ... 238
Escolher o Nosso Futuro ... 242
Modificar os Nossos Genes ... 245
Sair da Órbita de Darwin ... 248
Objetivo Número Um: Preservar o Conhecimento ... 248

Objetivo Número Dois: Adquirir Novo Conhecimento 251
Um Futuro com Propósito e Direção 251
Considerações Finais 255
Leituras Sugeridas 260
Agradecimentos 267
Acerca do Autor 270

Prefácio de RICHARD DAWKINS

Não leia este livro antes de dormir. Não porque seja assustador — não lhe dará pesadelos —, mas porque é tão estimulante, tão entusiasmante, que transformará a sua mente num redemoinho fervilhante de ideias provocadoras e inspiradoras — em vez de adormecer, vai querer correr a contar a alguém. Quem escreve este prefácio é uma vítima desse mesmo redemoinho, e creio que isso se notará.

Charles Darwin era um caso raro entre os cientistas, pois tinha os meios para trabalhar fora do ambiente universitário e sem depender de bolsas de investigação governamentais. Jeff Hawkins poderá não gostar de ser chamado o equivalente, no Vale do Silício, a um "cientista cavalheiro", mas — enfim, a analogia impõe-se. A ideia poderosa de Darwin era demasiado revolucionária para ser aceite quando expressa num breve artigo, e os documentos conjuntos Darwin-Wallace de 1858 foram praticamente ignorados. Como o próprio Darwin reconheceu, a ideia precisava de ser desenvolvida em formato de livro. E, de facto, foi o seu grande livro que abalou os alicerces da Era vitoriana, um ano depois. Também a Teoria dos Mil Cérebros de Jeff Hawkins requer um tratamento de fôlego — e igualmente a sua noção de "quadros de referência" (“O próprio ato de pensar é uma forma de movimento”) — certeiro como um tiro no alvo! Estas duas ideias são, cada uma por si, suficientemente profundas para preencherem um livro inteiro. Mas não ficamos por aqui.

T. H. Huxley disse, ao fechar *A Origem das Espécies:* “Que imensamente estúpido da minha parte não ter pensado nisto antes.” Não estou a sugerir que os neurocientistas digam o mesmo ao fechar este livro. Trata-se de um livro repleto de ideias fascinantes, mais do que de uma única ideia monumental como a de Darwin.

Suspeito que não apenas T. H. Huxley, mas também os seus três brilhantes netos, teriam adorado este livro: Andrew, porque descobriu como funciona o impulso nervoso (Hodgkin e Huxley são o Watson e Crick do sistema nervoso); Aldous, pelas suas viagens visionárias e poéticas aos confins da mente; e Julian, porque escreveu este poema, enaltecendo a capacidade do cérebro de construir um modelo da realidade, um microcosmo do universo:

O mundo das coisas entrou na tua mente infantil

Para povoar esse gabinete de cristal.
Dentro das suas paredes encontraram-se estranhos parceiros,
E coisas tornadas pensamento propagaram a sua espécie.
Pois, uma vez dentro, o facto corpóreo podia encontrar
Um espírito. Facto e tu, em dívida mútua,
Construíram ali o teu pequeno microcosmo — ao qual ainda
Foram atribuídas as mais vastas tarefas.
Homens mortos podem ali viver e conversar com estrelas:
O equador fala com o polo, e a noite com o dia;
O espírito dissolve as barreiras materiais do mundo —
Um milhão de isolamentos ardem até desaparecer.
O Universo pode viver e agir e planear,
Tornando-se enfim Deus na mente do homem.

O cérebro permanece na escuridão, apreendendo o mundo exterior apenas através de uma tempestade de impulsos nervosos — os impulsos de Andrew Huxley. Um impulso vindo do olho não é diferente de um vindo do ouvido ou do dedo grande do pé. O que os distingue é o local onde terminam no cérebro. Jeff Hawkins não é o primeiro cientista ou filósofo a sugerir que a realidade que percebemos é uma realidade construída — um modelo, constantemente atualizado e informado por boletins que nos chegam através dos sentidos. Mas Hawkins é, creio eu, o primeiro a dar espaço eloquente à ideia de que não existe apenas um modelo, mas milhares — um em cada uma das muitas colunas ordenadamente empilhadas que constituem o córtex cerebral. Existem cerca de 150.000 dessas colunas, e são elas as protagonistas da primeira parte do livro, juntamente com o que ele chama de "quadros de referência". A tese de Hawkins sobre ambos é provocadora, e será interessante ver como será recebida pelos restantes neurocientistas — creio que será bem acolhida. Uma das ideias mais fascinantes aqui é a de que as colunas corticais, nas suas atividades de modelação do mundo, funcionam de forma semiautónoma. Aquilo que "nós" percebemos é uma espécie de consenso democrático entre elas.

Democracia no cérebro? Consenso — e até disputa? Que ideia extraordinária. É um dos temas centrais do livro. Nós, mamíferos humanos, somos vítimas de uma disputa recorrente: um confronto entre o velho cérebro

reptiliano, que dirige inconscientemente a máquina da sobrevivência, e o neocórtex mamífero, que se senta, por assim dizer, numa espécie de assento do condutor, por cima dele. Este novo cérebro mamífero — o córtex cerebral — pensa. É o centro da consciência. Está ciente do passado, presente e futuro, e envia instruções ao velho cérebro, que as executa.

O velho cérebro, moldado pela seleção natural ao longo de milhões de anos, numa Era em que o açúcar era escasso e valioso para a sobrevivência, diz: “Bolo. Quero bolo. Hmmm, bolo. Dá-me.” O novo cérebro, moldado por livros e médicos durante apenas algumas dezenas de anos, numa Era em que o açúcar se tornou excessivo, diz: “Não, não. Nada de bolo. Não devo. Por favor, não comas esse bolo.” O velho cérebro diz: “Dor, dor, dor horrível, pára a dor imediatamente.” O novo cérebro diz: “Não, aguenta a tortura, não traias o teu país ao renderes-te. A lealdade à pátria e aos companheiros vem antes da tua própria vida.”

O conflito entre o velho cérebro reptiliano e o novo cérebro mamífero fornece a resposta para enigmas como: “Por que é que a dor tem de ser tão absurdamente dolorosa?” Afinal, para que serve a dor? A dor é uma substituta da morte. É um aviso ao cérebro: “Não faças isso outra vez — não provoques uma cobra, não apanhes uma brasa acesa, não saltes de grandes alturas. Desta vez só doeu; da próxima pode matar-te.” Mas agora um engenheiro de sistemas diria que o que precisávamos era de algo como uma bandeira indolor no cérebro. Quando essa bandeira se erguesse, seria o sinal de que não se deve repetir o comportamento. Em vez dessa solução fácil e indolor proposta pelo engenheiro, o que realmente recebemos é dor — muitas vezes excruciante, insuportável. Porquê? O que há de errado com a bandeira sensata?

A resposta talvez resida na natureza disputativa dos processos de tomada de decisão do cérebro: o conflito entre o velho cérebro e o novo. Se fosse demasiado fácil para o novo cérebro anular o voto do velho, o sistema da bandeira indolor não funcionaria. Nem mesmo a tortura funcionaria.

O novo cérebro sentir-se-ia livre para ignorar a minha bandeira hipotética e suportar um número indefinido de picadas de abelha, tornozelos torcidos ou parafusos dos inquisidores, se por algum motivo “quisesse”. O velho cérebro, que realmente “se importa” com a sobrevivência e com a transmissão dos genes, poderia “protestar” em vão. Talvez a seleção natural, em nome da sobrevivência, tenha garantido a “vitória” do velho cérebro tornando a dor tão absurdamente dolorosa que o novo cérebro não consiga sobrepor-se-lhe. Como outro exemplo: se o velho cérebro estivesse

“consciente” da traição ao propósito darwiniano do sexo, o simples ato de colocar um preservativo seria insuportavelmente doloroso.

Hawkins está do lado da maioria dos cientistas e filósofos informados que rejeitam totalmente o dualismo: não há nenhum “fantasma na máquina”, nenhuma alma misteriosa desligada do hardware que sobreviva à morte desse mesmo hardware, nenhum teatro cartesiano (expressão de Dan Dennett) onde um ecrã a cores exiba um filme do mundo a um “eu” que observa. Em vez disso, Hawkins propõe a existência de múltiplos modelos do mundo — microcosmos construídos e informados pelo fluxo constante de impulsos nervosos que jorra dos sentidos, sendo continuamente ajustados por ele. A propósito, Hawkins não exclui totalmente a possibilidade, num futuro longínquo, de escapar à morte através da transferência da mente para um computador, mas não acredita que isso venha a ser particularmente divertido.

Entre os modelos mais importantes que o cérebro constrói encontram-se os modelos do próprio corpo, os quais têm de lidar com o modo como os movimentos do corpo alteram a nossa perspetiva sobre o mundo exterior, além dos limites da prisão óssea do crânio. E isto é relevante para a grande preocupação da secção intermédia do livro: a inteligência das máquinas. Jeff Hawkins nutre grande respeito — tal como eu — por aquelas pessoas inteligentes, amigas dele e minhas, que temem a chegada de máquinas superinteligentes capazes de nos ultrapassar, subjugar, ou até eliminar por completo. Mas Hawkins não teme esse cenário, em parte porque as faculdades que tornam alguém exímio no xadrez ou no jogo de Go não são as mesmas que permitem lidar com a complexidade do mundo real. Crianças que não sabem jogar xadrez “sabem como se entornam líquidos, como rolam bolas e como ladram os cães. Sabem usar lápis, marcadores, papel e cola. Sabem abrir livros e que o papel pode rasgar-se.” E possuem uma autoimagem, uma imagem corporal que as posiciona no mundo físico e lhes permite navegar por ele sem esforço.

Não se trata de Hawkins subestimar o poder da inteligência artificial ou dos robôs do futuro. Muito pelo contrário. Mas ele considera que grande parte da investigação atual está a seguir o caminho errado. O caminho certo, na sua visão, passa por compreender como o cérebro funciona e imitar os seus mecanismos — mas acelerando-os drasticamente.

E não há razão para (na verdade, por favor, não o façamos) copiar os mecanismos do velho cérebro — os seus desejos e apetites, ânsias e fúrias, sentimentos e medos — que tantas vezes nos conduzem por caminhos que o novo cérebro considera prejudiciais. Prejudiciais, pelo menos, segundo a

perspetiva que Hawkins, eu e, muito provavelmente, você também, valorizamos. Porque ele é muito claro ao afirmar que os nossos valores iluminados devem, e de facto divergem radicalmente, do valor primário e primitivo dos nossos genes egoístas: o imperativo cru e cego de se reproduzir a todo o custo. Sem um velho cérebro, segundo a sua visão (que suspeito poderá ser controversa), não há razão para esperar que uma IA nutra sentimentos malévolos em relação a nós. Do mesmo modo — e também talvez de forma controversa — ele não considera que desligar uma IA consciente constitua um assassinato: sem um velho cérebro, por que é que ela sentiria medo ou tristeza? Por que haveria de querer sobreviver?

No capítulo "Genes versus Conhecimento", não resta qualquer dúvida quanto ao desfasamento entre os objetivos do velho cérebro (ao serviço dos genes egoístas) e os do novo cérebro (ao serviço do conhecimento). É a glória do córtex cerebral humano — único entre todos os animais e sem precedentes em toda a história geológica — possuir o poder de desafiar as ordens dos genes egoístas. Podemos usufruir do prazer sexual sem fins reprodutivos. Podemos dedicar as nossas vidas à filosofia, à matemática, à poesia, à astrofísica, à música, à geologia ou ao calor do amor humano, desafiando a insistência do velho cérebro (impulsionado pelos genes) de que tudo isso é uma perda de tempo — tempo que "deveria" ser empregue a combater rivais e a procurar múltiplos parceiros sexuais:

> Tal como o vejo, temos uma escolha profunda a fazer. É uma escolha entre favorecer o velho cérebro ou favorecer o novo cérebro. Mais concretamente, queremos que o nosso futuro seja impulsionado pelos processos que nos trouxeram até aqui — ou seja, a seleção natural, a competição e o impulso dos genes egoístas? Ou queremos que o nosso futuro seja conduzido pela inteligência e pelo seu desejo de compreender o mundo?

Comecei por citar a comovente e humilde observação de T. H. Huxley ao terminar a leitura de *A Origem das Espécies*, de Darwin. Terminarei agora com apenas uma das muitas ideias fascinantes de Jeff Hawkins — condensada em apenas duas páginas — que me levou a ecoar Huxley. Sentindo a necessidade de uma lápide cósmica, algo que permita à galáxia saber que estivemos aqui e que fomos capazes de o anunciar, Hawkins observa que todas as civilizações são efémeras. Na escala do tempo universal, o intervalo entre a invenção da comunicação eletromagnética por parte de uma civilização e a sua extinção é como o lampejo de um pirilampo. A probabilidade de esse lampejo coincidir com outro é, infelizmente, diminuta. O que é necessário então — e é por isso que Hawkins usa o termo

"lápide" — é uma mensagem que não diga apenas "Estamos aqui", mas sim "Estivemos aqui". E essa lápide tem de ter uma duração à escala cósmica: não só deve ser visível a vários parsecs de distância, como também tem de perdurar durante milhões — senão mesmo milhares de milhões — de anos, de forma a continuar a proclamar a sua mensagem quando, muito tempo depois da nossa extinção, outros lampejos de inteligência a possam eventualmente intercetar. Emitir números primos ou os dígitos de π não serve. Pelo menos não sob a forma de sinal de rádio ou de feixe laser pulsado. É certo que tais sinais proclamam a existência de inteligência biológica — razão pela qual são o recurso habitual do SETI (a procura de inteligência extraterrestre) e da ficção científica —, mas são demasiado breves, demasiado enraizados no presente. Que sinal, então, poderia durar o suficiente e ser detetável a grande distância, em qualquer direção? É aqui que Hawkins despertou o meu Huxley interior.

Está para além das nossas capacidades atuais, mas no futuro, antes de o nosso lampejo de pirilampo se extinguir, poderíamos colocar em órbita do Sol uma série de satélites "que bloqueassem parte da luz solar segundo um padrão que não poderia ocorrer naturalmente. Esses bloqueadores solares em órbita continuariam a rodear o Sol durante milhões de anos, muito depois de desaparecermos, e poderiam ser detetados a grande distância." Mesmo que o espaçamento entre esses satélites umbrais não corresponda literalmente a uma série de números primos, a mensagem poderia ser tornada inconfundível:

"Vida Inteligente Esteve Aqui" (Intelligent Life Woz 'Ere).

Algo que me agrada particularmente — e partilho este pequeno quadro com Jeff Hawkins como forma de lhe agradecer o prazer que o seu brilhante livro me proporcionou — é o facto de uma mensagem cósmica codificada na forma de um padrão de intervalos entre picos (ou, no seu caso, antipicos, uma vez que os seus satélites obscurecem o Sol) recorrer ao mesmo tipo de código que um neurónio utiliza.

Este é um livro sobre o modo como o cérebro funciona. E trabalha o cérebro de uma forma que não é menos do que exaltante.

PARTE 1

UMA NOVA COMPREENSÃO DO CÉREBRO

As células na sua cabeça estão a ler estas palavras. Pense em quão notável isso é. As células são simples. Uma única célula não sabe ler, nem pensar, nem fazer grande coisa. No entanto, se reunirmos células em número suficiente para formar um cérebro, elas não só leem livros — também os escrevem. Projetam edifícios, inventam tecnologias e decifram os mistérios do universo. Como é que um cérebro feito de células simples cria inteligência é uma questão profundamente interessante — e continua a ser um mistério.

Compreender como o cérebro funciona é considerado um dos grandes desafios da humanidade. Esta busca deu origem a dezenas de iniciativas nacionais e internacionais, como o Human Brain Project, na Europa, e a International Brain Initiative. Dezenas de milhares de neurocientistas trabalham em várias especialidades, em praticamente todos os países do mundo, tentando entender o cérebro. Embora os neurocientistas estudem os cérebros de diferentes animais e coloquem perguntas variadas, o objetivo último da neurociência é compreender como o cérebro humano dá origem à inteligência humana.

Poderá ficar surpreendido com a minha afirmação de que o cérebro humano continua a ser um mistério. Todos os anos são anunciadas novas descobertas relacionadas com o cérebro, são publicados novos livros sobre o tema, e investigadores de áreas relacionadas, como a inteligência artificial, afirmam que as suas criações estão a aproximar-se da inteligência, digamos, de um rato ou de um gato. Seria fácil concluir, com base nisso, que os cientistas têm já uma ideia bastante clara de como o cérebro funciona. Mas, se perguntar aos neurocientistas, quase todos admitirão que ainda estamos às escuras. Adquirimos uma enorme quantidade de

conhecimento e factos sobre o cérebro, mas temos muito pouca compreensão de como tudo isso funciona em conjunto.

Em 1979, Francis Crick, famoso pelo seu trabalho sobre o ADN, escreveu um ensaio sobre o estado da ciência do cérebro, intitulado *"Thinking About the Brain" (Pensar Sobre o Cérebro)*. Descrevia a grande quantidade de factos que os cientistas haviam reunido sobre o cérebro, mas concluía:

> Apesar da acumulação constante de conhecimento detalhado, o modo como o cérebro humano funciona continua a ser profundamente misterioso."

E acrescentava:

> "O que falta de forma gritante é um quadro amplo de ideias dentro do qual interpretar estes resultados.

Crick observou que os cientistas vinham a recolher dados sobre o cérebro há décadas. Sabiam imensos factos. Mas ninguém conseguia perceber como juntar esses factos de forma significativa. O cérebro era como um gigantesco puzzle com milhares de peças. As peças estavam diante de nós, mas não conseguíamos dar-lhes sentido. Ninguém sabia ao certo qual seria a imagem final. Segundo Crick, o cérebro era um mistério não por falta de dados, mas porque não sabíamos como organizar as peças que já tínhamos. Nos quarenta anos desde que Crick escreveu o seu ensaio, houve muitas descobertas significativas sobre o cérebro — várias das quais abordarei mais adiante —, mas, no essencial, a sua observação continua válida. O modo como a inteligência emerge das células da sua cabeça continua a ser um mistério profundo. À medida que se recolhem mais peças do puzzle a cada ano, por vezes tem-se a sensação de que nos estamos a afastar da compreensão do cérebro, em vez de nos aproximarmos.

Li o ensaio de Crick quando era jovem e ele inspirou-me. Senti que seria possível resolver o mistério do cérebro ainda durante a minha vida — e tenho perseguido esse objetivo desde então. Nos últimos quinze anos, tenho liderado uma equipa de investigação no Vale do Silício que estuda uma parte do cérebro chamada neocórtex. O neocórtex ocupa cerca de 70% do volume do cérebro humano e é responsável por tudo o que associamos à inteligência:

desde os nossos sentidos da visão, do tato e da audição, até à linguagem em todas as suas formas, passando pelo pensamento abstrato, como a matemática e a filosofia. O objetivo da nossa investigação é compreender o funcionamento do neocórtex com detalhe suficiente para que possamos explicar a biologia do cérebro e construir máquinas inteligentes que operem com base nos mesmos princípios.

No início de 2016, o progresso da nossa investigação mudou radicalmente. Tivemos um avanço decisivo na nossa compreensão. Percebemos que nós — e outros cientistas também — tínhamos ignorado um ingrediente fundamental. Com esse novo entendimento, começámos a ver como as peças do puzzle encaixavam. Ou seja, acredito que descobrimos o quadro de referência a que Crick se referia — um enquadramento que não só explica os fundamentos do funcionamento do neocórtex, como também dá origem a uma nova forma de pensar sobre a inteligência. Ainda não temos uma teoria completa do cérebro — estamos longe disso. As áreas científicas costumam começar com um enquadramento teórico, e só depois os detalhes vão sendo desenvolvidos. Talvez o exemplo mais famoso seja a teoria da evolução de Darwin: ele propôs uma nova forma arrojada de pensar sobre a origem das espécies, mas os detalhes, como o funcionamento dos genes e do ADN, só seriam conhecidos muitos anos mais tarde.

Para ser inteligente, o cérebro tem de aprender uma enorme quantidade de coisas sobre o mundo. Não me refiro apenas ao que aprendemos na escola, mas a coisas básicas, como o aspeto, o som e a sensação dos objetos do quotidiano. Temos de aprender como os objetos se comportam — desde como se abrem e fecham as portas até ao que acontece quando tocamos no ecrã das aplicações do nosso telemóvel. Precisamos de saber onde tudo está localizado no mundo, desde o lugar onde guardamos os nossos pertences em casa até à localização da biblioteca ou dos correios na nossa cidade. E, claro, aprendemos também conceitos de ordem superior, como o significado de "compaixão" ou "governo". Para além de tudo isto, cada um de nós aprende o significado de dezenas de milhares de palavras. Todos nós possuímos uma quantidade extraordinária de conhecimento sobre o mundo. Algumas das nossas capacidades

básicas são determinadas geneticamente — como comer ou recuar perante a dor —, mas a maioria do que sabemos sobre o mundo é adquirida.

Os cientistas dizem que o cérebro aprende um modelo do mundo. A palavra "modelo" implica que aquilo que sabemos não está armazenado como um simples amontoado de factos, mas organizado de uma forma que reflete a estrutura do mundo e tudo o que nele existe. Por exemplo, para sabermos o que é uma bicicleta, não memorizamos uma lista de factos sobre bicicletas. Em vez disso, o nosso cérebro cria um modelo de bicicleta que inclui as diferentes partes, como estas estão dispostas entre si e de que modo se movem e funcionam em conjunto. Para reconhecermos algo, temos de aprender primeiro como se parece e que sensação transmite; e, para atingirmos objetivos, temos de aprender como as coisas no mundo geralmente se comportam quando interagimos com elas. A inteligência está intimamente ligada ao modelo do mundo que o cérebro constrói; por isso, para compreendermos como o cérebro gera inteligência, temos de perceber como é que o cérebro, sendo composto por células simples, aprende um modelo do mundo e de tudo o que nele existe.

A nossa descoberta de 2016 explica como o cérebro aprende esse modelo. Deduziu-se que o neocórtex armazena tudo o que sabemos — todo o nosso conhecimento — utilizando algo a que chamámos estruturas de referência. Explicarei isto em mais detalhe mais adiante, mas, para já, considere-se o exemplo de um mapa em papel. Um mapa é um tipo de modelo: um mapa de uma cidade é um modelo dessa cidade, e as linhas de grelha, como as de latitude e longitude, constituem um tipo de estrutura de referência. As linhas do mapa, a sua grelha de coordenadas, fornecem a estrutura do modelo. Uma estrutura de referência diz-nos onde as coisas estão localizadas umas em relação às outras, e pode indicar-nos como alcançar objetivos — por exemplo, como ir de um lugar a outro. Percebemos que o modelo que o cérebro constrói do mundo é formado com base em estruturas de referência semelhantes às de um mapa. Não apenas uma estrutura, mas centenas de milhares delas. De facto, compreendemos agora que a maioria das células do neocórtex está dedicada à criação e manipulação de estruturas de referência, que o cérebro utiliza para planear e pensar.

Com esta nova descoberta, começaram a surgir respostas para algumas das maiores questões da neurociência. Perguntas como: Como é que os nossos diversos estímulos sensoriais se unificam numa experiência singular? O que acontece quando pensamos? Como podem duas pessoas chegar a crenças diferentes a partir das mesmas observações? E por que temos um sentido do "eu"?

Este livro conta a história dessas descobertas e as implicações que elas têm para o nosso futuro. A maior parte do material foi publicada em revistas científicas. No final do livro, forneço ligações para esses artigos. No entanto, os artigos científicos não são o meio mais adequado para explicar teorias de grande escala — sobretudo de uma forma acessível a não especialistas.

Dividi o livro em três partes. Na primeira parte, descrevo a nossa teoria das estruturas de referência, a que chamámos *Teoria dos Mil Cérebros*. Esta teoria baseia-se em parte na dedução lógica, por isso irei guiá-lo pelos passos que seguimos até chegarmos às nossas conclusões. Darei também algum contexto histórico, para que possa compreender como esta teoria se insere na longa história de reflexões sobre o cérebro. No final da primeira parte, espero que tenha uma compreensão clara do que se passa na sua mente quando pensa e age no mundo, e do que significa, afinal, ser inteligente.

A segunda parte do livro é dedicada à inteligência das máquinas. O século XXI será transformado pelas máquinas inteligentes, tal como o século XX foi transformado pelos computadores. A *Teoria dos Mil Cérebros* explica por que é que a inteligência artificial atual ainda não é verdadeiramente inteligente — e o que é necessário para criar máquinas realmente inteligentes. Descrevo como serão essas máquinas no futuro e de que forma poderemos utilizá-las. Explico por que é que algumas máquinas serão conscientes e o que, se algo, deveremos fazer em relação a isso. Por fim, muitas pessoas receiam que as máquinas inteligentes representem um risco existencial, temendo que estejamos prestes a criar uma tecnologia capaz de destruir a humanidade. Discordo. As nossas descobertas mostram por que é que a inteligência artificial, por si só, é inofensiva. Contudo, sendo uma tecnologia

poderosa, o risco reside na forma como os seres humanos a poderão utilizar.

Na terceira parte do livro, examino a condição humana a partir da perspetiva do cérebro e da inteligência. O modelo do mundo que o cérebro constrói inclui um modelo de nós próprios. Isto conduz à estranha verdade de que aquilo que você e eu percebemos, momento após momento, é uma simulação do mundo — e não o mundo real. Uma das consequências da *Teoria dos Mil Cérebros* é que as nossas crenças sobre o mundo podem ser falsas. Explico como isso pode acontecer, por que é que crenças falsas são difíceis de erradicar, e de que modo essas crenças, combinadas com as nossas emoções mais primitivas, representam uma ameaça à nossa sobrevivência a longo prazo.

Os capítulos finais abordam aquilo que considero ser a escolha mais importante que teremos de fazer enquanto espécie. Existem duas formas de pensarmos sobre nós próprios. Uma é como organismos biológicos, produtos da evolução e da seleção natural. Nesta visão, os seres humanos são definidos pelos seus genes, e o propósito da vida é replicá-los. Mas estamos agora a emergir do nosso passado puramente biológico. Tornámo-nos uma espécie inteligente. Somos a primeira espécie na Terra a conhecer a dimensão e a idade do universo. Somos a primeira a saber como a Terra evoluiu e como surgimos. Somos a primeira a desenvolver ferramentas que nos permitem explorar o universo e descobrir os seus segredos. Sob esta outra perspetiva, os seres humanos são definidos pela sua inteligência e conhecimento, não pelos seus genes. A escolha que se nos coloca para o futuro é: devemos continuar a ser guiados pelo nosso passado biológico ou, em vez disso, escolher abraçar a nossa inteligência emergente?

Talvez não possamos fazer as duas coisas. Estamos a criar tecnologias poderosas capazes de alterar radicalmente o planeta, manipular a biologia e, em breve, criar máquinas mais inteligentes do que nós. Mas continuamos a possuir os comportamentos primitivos que nos trouxeram até aqui. Esta combinação é o verdadeiro risco existencial que temos de enfrentar. Se estivermos dispostos a assumir a inteligência e o conhecimento como o que nos

define — em vez dos nossos genes — então talvez consigamos criar um futuro mais duradouro e com um propósito mais nobre.

A jornada que conduziu à *Teoria dos Mil Cérebros* foi longa e sinuosa. Estudei engenharia eletrotécnica na universidade e tinha acabado de começar o meu primeiro emprego na Intel quando li um ensaio de Francis Crick. Esse ensaio teve um impacto tão profundo em mim que decidi mudar de carreira e dedicar a minha vida ao estudo do cérebro. Depois de uma tentativa falhada de conseguir um lugar para estudar o cérebro na própria Intel, candidatei-me ao programa de pós-graduação do laboratório de Inteligência Artificial do MIT. (Sentia que a melhor forma de construir máquinas inteligentes era estudar primeiro o cérebro.) Durante as entrevistas com os professores do MIT, a minha proposta de criar máquinas inteligentes baseadas em teorias do cérebro foi rejeitada. Disseram-me que o cérebro era apenas um computador desorganizado e que não valia a pena estudá-lo. Desencorajado, mas não derrotado, inscrevi-me então num programa de doutoramento em neurociência na Universidade da Califórnia, em Berkeley. Iniciei os meus estudos em janeiro de 1986.

Ao chegar a Berkeley, contatei o responsável pelo grupo de pós-graduação em neurobiologia, o Dr. Frank Werblin, para pedir aconselhamento. Ele pediu-me que escrevesse um artigo a descrever a investigação que queria realizar para a minha tese de doutoramento. Nesse artigo, expliquei que pretendia trabalhar numa teoria do neocortex. Sabia que queria abordar o problema estudando como o neocortex faz previsões. O professor Werblin fez com que vários membros do corpo docente lessem o meu artigo, que foi bem recebido. Ele disse-me que as minhas ambições eram admiráveis, que a minha abordagem era sólida e que o problema que queria abordar era um dos mais importantes da ciência, mas — e eu não esperava por isto — ele não via como eu poderia realizar o meu sonho naquele momento. Enquanto estudante de neurociência, teria de trabalhar para um professor, desenvolvendo um trabalho semelhante ao que o professor já fazia. E ninguém em Berkeley, nem em qualquer outro lugar que ele conhecesse, estava a fazer algo suficientemente próximo do que eu queria fazer.

Tentar desenvolver uma teoria global da função cerebral era considerado demasiado ambicioso e, por isso, demasiado arriscado. Se um estudante trabalhasse neste tema durante cinco anos e não apresentasse progresso, poderia não se formar. Para os professores, o risco era semelhante; poderiam não obter estabilidade académica. As agências que financiavam a investigação também consideravam o tema demasiado arriscado. Propostas de investigação focadas em teoria eram rotineiramente rejeitadas.

Poderia ter trabalhado num laboratório experimental, mas, após entrevistas em alguns deles, percebi que não era o ideal para mim. Passaria a maior parte do tempo a treinar animais, construir equipamentos experimentais e recolher dados. As teorias que desenvolvesse seriam limitadas à parte do cérebro estudada nesse laboratório.

Durante os dois anos seguintes, passei os meus dias nas bibliotecas da universidade a ler artigo atrás de artigo de neurociência. Li centenas, incluindo todos os artigos mais importantes publicados nos cinquenta anos anteriores. Também li o que psicólogos, linguistas, matemáticos e filósofos pensavam sobre o cérebro e a inteligência. Obtive uma educação de primeira classe, embora não convencional. Depois de dois anos de estudo autodidata, senti que precisava de uma mudança. Elaborei um plano: iria trabalhar novamente na indústria durante quatro anos e depois reavaliaria as minhas oportunidades no meio académico. Assim, voltei a trabalhar com computadores pessoais no Vale do Silício.

Comecei a ter sucesso como empreendedor. De 1988 a 1992, criei um dos primeiros computadores tablet, o GridPad. Depois, em 1992, fundei a Palm Computing, iniciando um período de dez anos durante o qual desenhei alguns dos primeiros computadores de mão e smartphones, como o PalmPilot e o Treo. Todos os que trabalhavam comigo na Palm sabiam que o meu coração estava na neurociência, que via o meu trabalho em computação móvel como algo temporário. Conceber alguns dos primeiros computadores portáteis e smartphones era um trabalho entusiasmante. Sabia que milhares de milhões de pessoas acabariam por depender destes

dispositivos, mas sentia que compreender o cérebro era ainda mais importante. Acreditava que uma teoria cerebral teria um impacto positivo maior no futuro da humanidade do que a computação. Por isso, precisava de regressar à investigação cerebral.

Não existia um momento conveniente para partir, por isso escolhi uma data e afastei-me dos negócios que ajudei a criar. Com a ajuda e incentivo de alguns amigos neurocientistas (notadamente Bob Knight na UC Berkeley, Bruno Olshausen na UC Davis e Steve Zornetzer na NASA Ames Research), criei, em 2002, o Redwood Neuroscience Institute (RNI). O RNI focava-se exclusivamente na teoria do neocortex e tinha dez cientistas a tempo inteiro. Todos nós estávamos interessados em teorias de larga escala do cérebro, e o RNI era um dos poucos lugares no mundo onde este foco não só era tolerado, como esperado. Ao longo dos três anos em que dirigi o RNI, recebemos mais de cem investigadores visitantes, alguns dos quais ficaram dias ou semanas. Tivemos palestras semanais, abertas ao público, que normalmente se prolongavam por horas de discussão e debate.

Todos os que trabalhavam no RNI, incluindo eu, achavam que era um ambiente excelente. Tive oportunidade de conhecer e conviver com muitos dos melhores neurocientistas do mundo. Isso permitiu-me adquirir conhecimento em múltiplos campos da neurociência, algo difícil de conseguir numa posição académica típica. O problema era que eu queria conhecer as respostas a um conjunto específico de questões, e não via a equipa a mover-se no sentido de um consenso sobre essas questões. Os cientistas individuais estavam satisfeitos por seguir os seus próprios caminhos. Por isso, após três anos a dirigir um instituto, decidi que a melhor forma de alcançar os meus objetivos seria liderar a minha própria equipa de investigação.

O RNI estava a prosperar em todos os outros aspetos, por isso decidimos transferi-lo para a UC Berkeley. Sim, o mesmo lugar que me disse que não podia estudar teoria cerebral decidiu, dezenove anos depois, que um centro de teoria cerebral era exatamente o que eles precisavam. O RNI continua hoje como o Redwood Center for Theoretical Neuroscience.

Com a mudança do RNI para a UC Berkeley, vários colegas e eu fundámos a Numenta. A Numenta é uma empresa de investigação independente. O nosso objetivo principal é desenvolver uma teoria sobre o funcionamento do neocortex. O nosso objetivo secundário é aplicar o que aprendemos sobre o cérebro à aprendizagem automática e à inteligência das máquinas. A Numenta é semelhante a um laboratório de investigação típico numa universidade, mas com maior flexibilidade. Isso permite-me liderar uma equipa, assegurar que estamos todos focados na mesma tarefa e experimentar novas ideias sempre que necessário.

Enquanto escrevo, a Numenta tem mais de quinze anos, mas, em certos aspetos, ainda somos como uma start-up. Tentar perceber como o neocortex funciona é extremamente desafiante. Para progredir, precisamos da flexibilidade e do foco que um ambiente de start-up proporciona. Também precisamos de muita paciência, algo que não é típico numa start-up. A nossa primeira descoberta significativa — como os neurónios fazem previsões — ocorreu em 2010, cinco anos depois de termos começado. A descoberta dos quadros de referência tipo mapas no neocortex ocorreu seis anos mais tarde, em 2016.

Em 2019, começámos a trabalhar na nossa segunda missão, que é aplicar os princípios do cérebro à aprendizagem automática. Foi também nesse ano que comecei a escrever este livro, para partilhar aquilo que aprendemos.

Acho incrível que a única coisa no universo que sabe que o universo existe seja essa massa de três libras de células flutuando nas nossas cabeças. Isso lembra-me do velho enigma: Se uma árvore cai na floresta e ninguém está lá para a ouvir, terá feito algum som? De modo semelhante, podemos perguntar: Se o universo entrou e saiu da existência e não houvesse cérebros para o saber, terá o universo realmente existido? Quem o saberia? Alguns milhares de milhões de células suspensas no seu crânio sabem não só que o universo existe, mas que é vasto e antigo.

Estas células aprenderam um modelo do mundo, um conhecimento que, tanto quanto sabemos, não existe em mais lado nenhum. Dediquei a minha vida a tentar compreender como o

cérebro faz isto, e sinto-me entusiasmado com o que aprendemos. Espero que você também esteja entusiasmado. Vamos começar.

CAPÍTULO 1

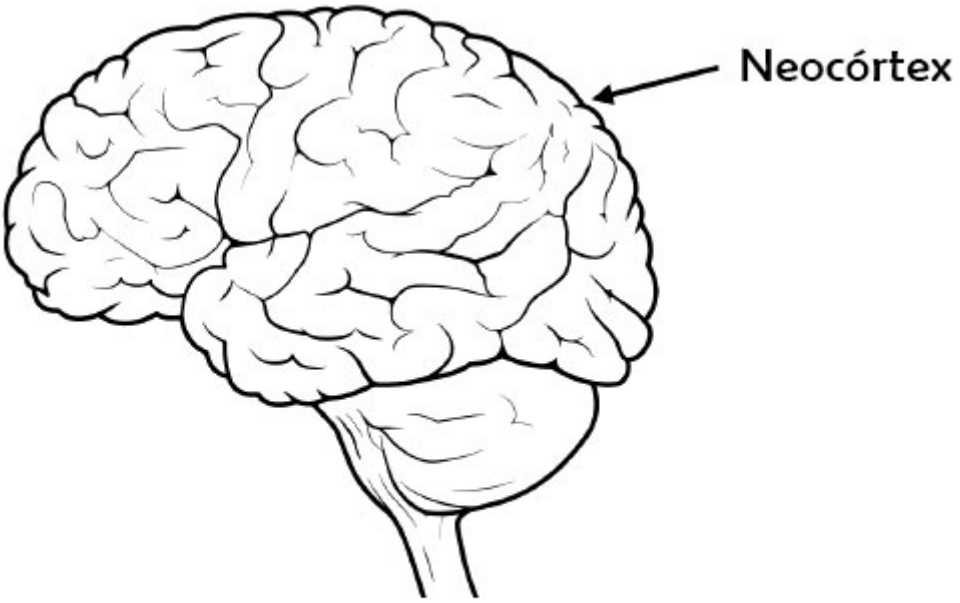
Cérebro Velho – Cérebro Novo

Para compreender como o cérebro cria inteligência, há alguns fundamentos que é necessário conhecer primeiro.

Pouco depois de Charles Darwin publicar a sua teoria da evolução, os biólogos perceberam que o próprio cérebro humano tinha evoluído ao longo do tempo — e que a sua história evolutiva é visível apenas ao olharmos para ele. Ao contrário de muitas espécies, que frequentemente desaparecem à medida que novas surgem, o cérebro evoluiu por adição: acrescentando novas partes sobre as mais antigas. Por exemplo, alguns dos sistemas nervosos mais antigos e simples consistem em conjuntos de neurónios que percorrem as costas de pequenos vermes. Estes neurónios permitem ao verme realizar movimentos simples, sendo os antecessores da nossa medula espinal, que é igualmente responsável por muitos dos nossos movimentos básicos. A seguir, surgiu uma massa de neurónios numa das extremidades do corpo, que passou a controlar funções como a digestão e a respiração. Essa massa é o antecessor do nosso tronco cerebral, que, de forma análoga, controla a digestão e a respiração. O tronco cerebral veio estender aquilo que já existia, mas não o substituiu. Com o tempo, o cérebro foi adquirindo capacidade para comportamentos cada vez mais complexos, à medida que evoluíam novas partes sobre as anteriores. Este processo de crescimento por acumulação aplica-se aos cérebros da maioria dos animais complexos. E é fácil perceber por que é que as partes antigas do cérebro ainda lá estão: por mais inteligentes e sofisticados que sejamos, respirar, comer, ter relações sexuais e reagir por reflexo continuam a ser funções vitais para a nossa sobrevivência.

A parte mais recente do nosso cérebro é o neocórtex, que significa “nova camada externa”. Todos os mamíferos — e apenas os mamíferos — possuem um neocórtex. O neocórtex humano é particularmente grande, ocupando cerca de 70% do volume total

do cérebro. Se conseguíssemos retirar o neocórtex da cabeça e estendê-lo numa superfície plana, teria aproximadamente o tamanho de um grande guardanapo de mesa e o dobro da sua espessura (cerca de 2,5 milímetros). Ele envolve as partes mais antigas do cérebro de tal forma que, quando olhamos para um cérebro humano, quase tudo o que vemos é o neocórtex (com as suas características pregas e sulcos), com pequenas porções do cérebro antigo e da medula espinal a emergirem pela base.



Um cérebro humano

O neocórtex é o órgão da inteligência. Quase todas as capacidades que associamos à inteligência — como a visão, a linguagem, a música, a matemática, a ciência ou a engenharia — são criadas pelo neocórtex. Quando pensamos em algo, é sobretudo o neocórtex que está a "pensar". É o seu neocórtex que está a ler ou a ouvir este livro, e é o meu neocórtex que o está a escrever. Se quisermos compreender a inteligência, então temos de compreender o que faz o neocórtex — e como o faz.

Um animal não precisa de neocórtex para viver uma vida complexa. O cérebro de um crocodilo é, em termos gerais, equivalente ao nosso, mas sem um neocórtex propriamente dito. Ainda assim, o crocodilo exibe comportamentos sofisticados, cuida das suas crias e sabe como navegar pelo ambiente. A maioria das

pessoas concordaria que o crocodilo possui algum grau de inteligência — embora nada que se compare à inteligência humana.

O neocórtex e as partes mais antigas do cérebro estão interligados por fibras nervosas; por isso, não podemos considerá-los órgãos completamente separados. São mais como colegas de casa: têm agendas e "personalidades" diferentes, mas precisam de cooperar para que algo seja feito. O neocórtex encontra-se, no entanto, numa posição decididamente injusta, pois não controla o comportamento de forma direta. Ao contrário de outras partes do cérebro, nenhuma das suas células se liga diretamente aos músculos, o que significa que não pode, por si só, provocar qualquer movimento. Quando o neocórtex quer fazer algo, envia um sinal ao cérebro antigo — pedindo-lhe, por assim dizer, que execute a sua vontade. Por exemplo, respirar é uma função do tronco cerebral, e não requer qualquer pensamento ou intervenção do neocórtex. É possível controlar a respiração temporariamente — como quando decidimos prender o fôlego —, mas se o tronco cerebral detetar que o corpo precisa de mais oxigénio, retoma o controlo e ignora o neocórtex. De modo semelhante, o neocórtex pode pensar: "Não comas este bolo, não é saudável." Mas se as partes mais antigas e primitivas do cérebro disserem: "Parece bom, cheira bem — come", o bolo pode tornar-se difícil de resistir. Este conflito entre o cérebro antigo e o cérebro novo é um dos temas de fundo deste livro. Terá um papel crucial quando abordarmos os riscos existenciais que a humanidade enfrenta.

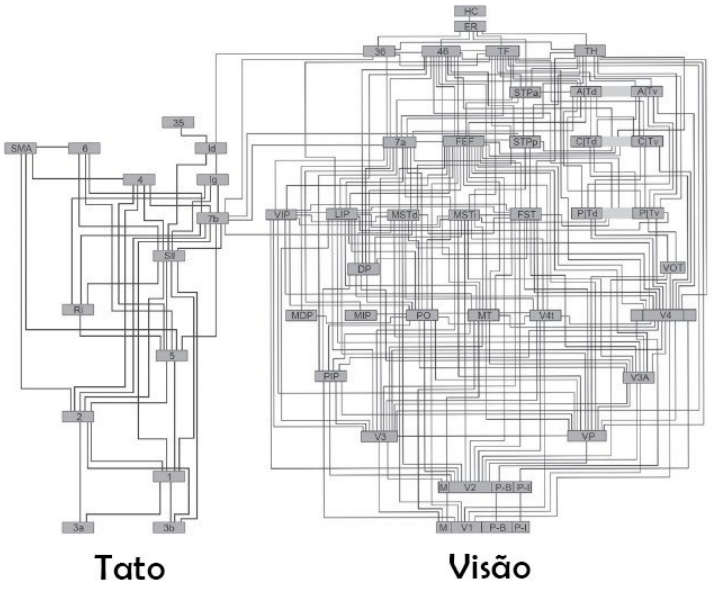
O cérebro antigo contém dezenas de órgãos distintos, cada um com uma função específica. São visivelmente diferenciáveis, e as suas formas, dimensões e ligações refletem o que fazem. Por exemplo, existem vários pequenos órgãos do tamanho de uma ervilha na amígdala, uma região antiga do cérebro, que são responsáveis por diferentes tipos de agressividade — como a agressividade premeditada ou a impulsiva.

O neocórtex é surpreendentemente diferente. Apesar de ocupar cerca de três quartos do volume cerebral e ser responsável por inúmeras funções cognitivas, não apresenta divisões visualmente óbvias. As suas pregas e sulcos existem para que o neocórtex caiba no crânio — algo semelhante ao que aconteceria se

se tentasse enfiar um guardanapo num copo largo de vinho. Se ignorarmos as pregas, o neocórtex assemelha-se a uma única folha de células, sem divisões evidentes.

Ainda assim, o neocórtex está funcionalmente dividido em várias dezenas de áreas ou regiões, cada uma responsável por diferentes tarefas. Algumas tratam da visão, outras da audição ou do tato. Existem também regiões responsáveis pela linguagem e pelo planeamento. Quando o neocórtex sofre danos, as deficiências resultantes dependem da zona afetada. Danos na parte posterior da cabeça podem causar cegueira, e danos no lado esquerdo podem levar à perda da linguagem.

As várias regiões do neocórtex estão interligadas por feixes de fibras nervosas que circulam por baixo dele — a chamada "matéria branca" do cérebro. Ao seguir cuidadosamente essas fibras, os cientistas conseguem determinar quantas regiões existem e como se encontram conetadas. Estudar cérebros humanos é difícil, pelo que o primeiro mamífero complexo a ser analisado dessa forma foi o macaco-de-cauda-curta (macaco-macaque). Em 1991, dois cientistas, Daniel Felleman e David Van Essen, combinaram dados de dezenas de estudos para criar uma ilustração célebre do neocórtex do macaco-macaque. (Um mapa do neocórtex humano teria pormenores diferentes, mas uma estrutura geral semelhante.)



Tato Visão

Ligações no neocórtex

Os vários pequenos retângulos nesta imagem representam as diferentes regiões do neocórtex, e as linhas representam o modo como a informação flui de uma região para outra através da matéria branca.

Uma interpretação comum desta imagem é que o neocórtex funciona de forma hierárquica, como um organograma. A informação sensorial entra na base (neste diagrama, os estímulos provenientes da pele surgem à esquerda e os estímulos visuais, à direita). Esses dados sensoriais são processados em várias etapas sucessivas, em que cada região extrai características cada vez mais complexas do input inicial. Por exemplo, a primeira região a receber sinais dos olhos pode detetar padrões simples, como linhas ou contornos. Esta informação é depois transmitida à próxima região, que pode reconhecer elementos mais complexos, como ângulos ou formas. Este processo em etapas continua até que algumas regiões sejam capazes de reconhecer objetos completos.

Há várias provas que sustentam esta interpretação hierárquica. Por exemplo, quando os cientistas observam células em regiões situadas na base da hierarquia, verificam que estas reagem sobretudo a características simples, enquanto as células em regiões seguintes respondem a padrões mais complexos. Por vezes, encontram-se células em zonas superiores que respondem a objetos completos. Contudo, também há muita evidência a indicar que o neocórtex não se comporta como um organograma. Como se pode observar no diagrama, as regiões não estão organizadas verticalmente, umas sobre as outras, como num modelo hierárquico clássico. Existem várias regiões em cada nível, e a maioria das regiões está interligada a múltiplos níveis da hierarquia. Na verdade, a maioria das conexões entre regiões não segue um esquema hierárquico. Além disso, apenas algumas das células em cada região atuam como detetores de características; os cientistas ainda não conseguiram determinar o que faz a maioria das células em cada região.

Ficamos, assim, perante um enigma. O órgão da inteligência — o neocórtex — está dividido em dezenas de regiões com funções distintas, mas que, à superfície, parecem todas iguais. As suas ligações formam uma rede intricada, algo semelhante a um

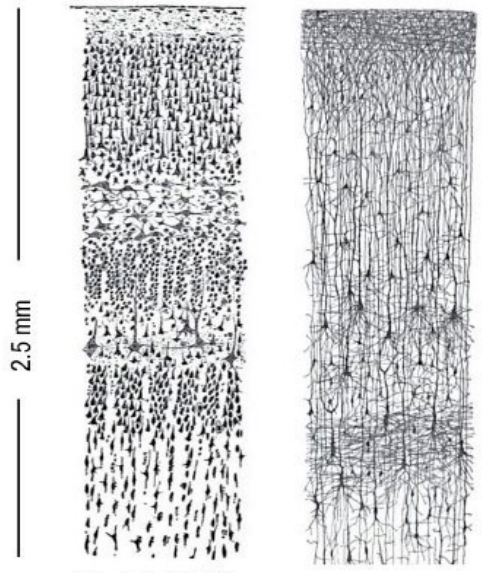
organograma, mas na verdade muito diferente. Não é de imediato evidente por que razão o órgão da inteligência tem esta aparência.

O passo seguinte parece óbvio: olhar para dentro do neocórtex e examinar os circuitos neuronais detalhados no seu interior, ao longo dos seus 2,5 milímetros de espessura. Poderíamos imaginar que, mesmo que as diferentes áreas do neocórtex se assemelhem por fora, os circuitos neuronais que criam funções como a visão, o tato ou a linguagem teriam de ser diferentes no seu interior. Mas isso não se confirma.

O primeiro a observar o circuito detalhado no interior do neocórtex foi Santiago Ramón y Cajal. No final do século XIX, foram descobertas técnicas de coloração que permitiam visualizar neurónios individuais ao microscópio. Cajal utilizou essas colorações para realizar ilustrações de todas as partes do cérebro. Criou milhares de imagens que, pela primeira vez, revelavam o aspeto do cérebro ao nível celular. Todas as representações belas e minuciosas que Cajal fez do cérebro foram desenhadas à mão. O seu trabalho pioneiro valeu-lhe, mais tarde, o Prémio Nobel. Entre as muitas imagens produzidas, há duas em particular que representam o neocórtex.

- A imagem da esquerda mostra apenas os corpos celulares dos neurónios.
- A da direita inclui também as ligações entre as células.

Estas imagens ilustram um corte transversal através dos 2,5 milímetros de espessura do neocórtex — permitindo ver, em detalhe, como os neurónios estão organizados e conetados no interior desta camada fundamental do cérebro humano.



Neurónios numa fatia do neocórtex

As colorações utilizadas para criar estas imagens tingem apenas uma pequena percentagem das células. E ainda bem — pois, se todas as células fossem coradas, veríamos apenas uma mancha negra. É importante lembrar que o número real de neurónios é muito maior do que aquilo que aparece nestas imagens.

A primeira coisa que Cajal e outros observaram foi que os neurónios no neocórtex parecem estar organizados em camadas. Estas camadas, que se estendem paralelamente à superfície do neocórtex (ou seja, na horizontal da imagem), resultam de diferenças no tamanho dos neurónios e na densidade com que estão dispostos. Imagine, por exemplo, um tubo de vidro onde se colocam uma camada de ervilhas, outra de lentilhas e uma terceira de grãos de soja. Visto de lado, observar-se-iam três camadas distintas. O mesmo princípio aplica-se às imagens acima. Quantas camadas existem depende de quem observa e dos critérios utilizados. Cajal identificou seis camadas. Uma interpretação simples é a de que cada camada de neurónios desempenha uma função diferente.

Hoje sabemos que o neocórtex contém dezenas de tipos distintos de neurónios, não apenas seis. Ainda assim, os cientistas continuam a usar a terminologia das seis camadas. Por exemplo, um determinado tipo de célula pode ser encontrado na Camada 3, enquanto outro tipo está mais presente na Camada 5.

- A Camada 1 é a mais superficial, próxima do crânio, no topo do desenho de Cajal.
- A Camada 6 é a mais profunda, próxima do centro do cérebro.

Convém sublinhar que estas camadas são apenas um guia aproximado: o mais relevante é com que outras células o neurónio se liga e como se comporta. Quando se classificam os neurónios com base nas suas conetividades, identificam-se dezenas de tipos distintos.

A segunda observação resultante destas imagens é que a maioria das conexões entre neurónios corre na vertical, entre camadas. Os neurónios possuem extensões ramificadas chamadas axónios e dendrites, que lhes permitem enviar e receber informação. Cajal observou que a maioria dos axónios se orienta entre as camadas, de forma perpendicular à superfície do neocórtex (ou seja, de cima para baixo nas imagens). Há neurónios em algumas camadas que estabelecem ligações horizontais de longa distância, mas a maioria das conexões são verticais. Isso significa que a informação que chega a uma região do neocórtex circula sobretudo de forma ascendente e descendente entre camadas, antes de ser enviada para outras partes do cérebro.

Nos 120 anos desde que Cajal fez estas primeiras imagens do cérebro, centenas de cientistas estudaram o neocórtex para desvendar todos os detalhes possíveis dos seus neurónios e circuitos. Existem hoje milhares de artigos científicos sobre este tema — demasiados para serem resumidos aqui. Por isso, o autor opta por destacar três observações gerais, que apresentará a seguir.

1. Os Circuitos Locais no Neocórtex São Complexos

Num único milímetro quadrado de neocórtex (aproximadamente 2,5 milímetros cúbicos), existem cerca de cem mil neurónios, quinhentas milhões de conexões sinápticas (ou seja, ligações entre neurónios), e vários quilómetros de axónios e dendrites. Imagine estender vários quilómetros de fio ao longo de uma estrada, e depois tentar comprimir tudo isso no volume de um grão de arroz — esta é a densidade e complexidade do circuito

que encontramos numa ínfima porção de neocórtex. Sob cada milímetro quadrado existem dezenas de tipos diferentes de neurónios. Cada tipo estabelece ligações específicas (prototípicas) com outros tipos neuronais. Muitas vezes os cientistas descrevem certas regiões do neocórtex como sendo responsáveis por funções simples, como a deteção de padrões ou de características. Contudo, bastariam apenas alguns neurónios para esse tipo de tarefa. A existência de circuitos neuronais tão precisos e extraordinariamente complexos em todas as regiões do neocórtex indica que cada uma está a realizar algo muito mais elaborado do que simplesmente detetar padrões.

2. O Neocórtex Apresenta um Aspeto Semelhante em Todo o Lado

Apesar da complexidade, o circuito do neocórtex tem um aspeto surpreendentemente uniforme em regiões visuais, linguísticas, táteis e outras. Este padrão repetido também se verifica entre espécies distintas, como ratos, gatos e humanos. Claro que há diferenças subtis: por exemplo, certas regiões contêm maior abundância de determinados tipos de células e menos de outros; há ainda regiões com tipos celulares únicos, que não se encontram noutras zonas do cérebro. Presume-se que estas variações refletem adaptações funcionais específicas. Ainda assim, as diferenças são relativamente pequenas quando comparadas com as semelhanças estruturais entre regiões. Esta uniformidade sugere que o neocórtex segue um princípio organizador comum, independentemente da função que desempenha numa dada região.

3. Todas as Partes do Neocórtex Geram Movimento

Durante muito tempo, acreditou-se que a informação entrava no neocórtex pelas chamadas "regiões sensoriais", ascendia e descia ao longo de uma hierarquia de regiões, até finalmente descer para a "região motora", cujas células projetam para os neurónios da espinal medula que movem os músculos e os membros. Hoje sabemos que esta

descrição é enganadora. Em todas as regiões examinadas, os cientistas encontraram células que projetam para partes do cérebro antigo relacionadas com o movimento. Por exemplo:

- As regiões visuais, que recebem informação dos olhos, enviam sinais para a parte do cérebro antigo que controla o movimento ocular.
- As regiões auditivas, que recebem informação dos ouvidos, projetam para áreas que controlam o movimento da cabeça.

Mover a cabeça altera o que se ouve, tal como mover os olhos altera o que se vê. A evidência de que dispomos indica que a complexidade dos circuitos do neocórtex tem sempre uma componente sensório-motora. Não existem regiões puramente motoras nem puramente sensoriais.

Em resumo, o neocórtex é o órgão da inteligência. Trata-se de uma folha de tecido neural, com o tamanho aproximado de um guardanapo de jantar, dividida em dezenas de regiões. Há regiões especializadas na visão, audição, tato e linguagem. Outras, de função menos óbvia, estão envolvidas no pensamento de ordem superior e no planeamento. Essas regiões estão interligadas através de feixes de fibras nervosas. Algumas das conexões seguem uma lógica hierárquica, parecendo organizar o fluxo de informação de modo sequencial, como num organograma. Outras, porém, revelam pouca ou nenhuma ordem aparente, sugerindo que a informação circula livremente entre múltiplas regiões em simultâneo. Independentemente da função que desempenham, todas as regiões do neocórtex partilham uma estrutura interna surpreendentemente semelhante.

Na próxima parte, conheceremos a primeira pessoa que conseguiu dar sentido a todas estas observações.

Este é um bom momento para dizer algumas palavras sobre o estilo de escrita deste livro. Estou a escrever para um leitor leigo, mas intelectualmente curioso. O meu objetivo é transmitir tudo o que precisa de saber para compreender a nova teoria — mas não muito mais do que isso. Parto do princípio de que a maioria dos leitores terá conhecimentos limitados de neurociência. No entanto, se tiver formação na área, reconhecerá onde omito pormenores e simplifico tópicos complexos. Se for esse o seu caso, peço-lhe compreensão. No final do livro encontrará uma lista de leitura anotada, onde explico onde poderá encontrar mais pormenores, caso tenha interesse em aprofundar.

CAPÍTULO 2

A Grande Ideia de Vernon Mountcastle

O cérebro atento é um livro pequeno, com apenas cem páginas. Publicado em 1978, contém dois ensaios sobre o cérebro, escritos por dois cientistas proeminentes. Um desses ensaios, de Vernon Mountcastle, neurocientista da Universidade Johns Hopkins, continua a ser uma das monografias mais icónicas e importantes alguma vez escritas sobre o cérebro. Mountcastle propôs uma forma de pensar o cérebro que é elegante — uma característica das grandes teorias — mas também tão surpreendente que continua a polarizar a comunidade científica da neurociência.

Li pela primeira vez *O Cérebro Atento* em 1982. O ensaio de Mountcastle teve um efeito imediato e profundo em mim e, como verá, a sua proposta influenciou fortemente a teoria que apresento neste livro.

A escrita de Mountcastle é precisa e erudita, mas também desafiante de ler. O título do seu ensaio é pouco apelativo: "Um Princípio Organizador para a Função Cerebral: A Unidade Módulo e o Sistema Distribuído". As linhas iniciais são difíceis de compreender; incluo-as aqui para que tenha uma ideia do tom do seu ensaio:

> **Não há dúvida da influência dominante da revolução darwiniana do século XIX nos conceitos sobre a estrutura e função do sistema nervoso. As ideias de Spencer, Jackson, Sherrington e de muitos outros que os seguiram assentavam na teoria evolutiva de que o cérebro se desenvolve na filogenia pela adição sucessiva de partes mais cefálicas. Segundo esta teoria, cada nova adição ou ampliação era acompanhada pela elaboração de comportamentos mais**

> **complexos e, ao mesmo tempo, impunha uma regulação às partes mais caudais e primitivas e ao comportamento presumivelmente mais primitivo que controlavam.**

O que Mountcastle explica nestas três primeiras frases é que o cérebro cresceu ao longo da evolução ao adicionar novas partes cerebrais sobre as partes mais antigas. As partes mais antigas controlam comportamentos mais primitivos, enquanto as partes mais recentes criam comportamentos mais sofisticados. Espero que isto lhe seja familiar, pois já falei desta ideia no capítulo anterior.

Contudo, Mountcastle avança que, embora grande parte do cérebro tenha crescido pela adição de novas partes sobre as antigas, não foi assim que o neocórtex cresceu para ocupar 70% do nosso cérebro. O neocórtex cresceu fazendo muitas cópias do mesmo circuito básico. Imagine assistir a um vídeo da evolução do nosso cérebro. O cérebro começa pequeno. Aparece uma nova parte numa extremidade, depois outra por cima, e depois outra por cima dessas. Em determinado momento, há milhões de anos, aparece uma nova parte que agora chamamos de neocórtex. O neocórtex começa pequeno, mas cresce não criando algo novo, mas copiando o circuito básico repetidamente. À medida que o neocórtex cresce, aumenta em área, mas não em espessura. Mountcastle argumentou que, embora o neocórtex humano seja muito maior do que o do rato ou do cão, todos são feitos do mesmo elemento — apenas temos mais cópias desse elemento.

O ensaio de Mountcastle recorda-me o livro de Charles Darwin, *A Origem das Espécies*. Darwin estava nervoso que a sua teoria da evolução causasse polémica. Por isso, no livro, ele aborda muitos conteúdos densos e relativamente desinteressantes sobre a variação no reino animal antes de descrever finalmente a sua teoria, já perto do fim. Mesmo assim, nunca afirma explicitamente que a evolução se aplica aos humanos. Ao ler o ensaio de Mountcastle, tenho uma impressão semelhante. Parece que Mountcastle sabe que a sua proposta vai gerar resistência, por isso escreve de forma cuidadosa e deliberada. Eis uma segunda citação, retirada de uma parte posterior do ensaio de Mountcastle:

> **Resumindo, não há nada intrinsecamente motor no córtex motor, nem sensorial no córtex sensorial. Assim, a elucidação do modo de funcionamento do circuito modular local em qualquer parte do neocórtex terá um significado generalizador muito grande.**

Nestas duas frases, Mountcastle resume a ideia principal do seu ensaio. Ele diz que cada parte do neocórtex funciona segundo o mesmo princípio. Todas as capacidades que associamos à inteligência — desde a visão, ao tato, à linguagem, ao pensamento de alto nível — são fundamentalmente a mesma coisa.

Recorde que o neocórtex está dividido em dezenas de regiões, cada uma com uma função diferente. Mas, olhando para o neocórtex de fora, não se veem essas regiões; não há demarcações, tal como numa imagem de satélite não se revelam fronteiras políticas entre países. Se cortarmos o neocórtex, vemos uma arquitetura complexa e detalhada, mas os detalhes são semelhantes, independentemente da região cortada. Uma fatia do córtex responsável pela visão parece igual a uma fatia do córtex responsável pelo tato, que parece igual a uma fatia do córtex responsável pela linguagem.

Mountcastle propôs que a razão pela qual as regiões parecem semelhantes é porque todas fazem a mesma coisa. O que as distingue não é a sua função intrínseca, mas aquilo a que estão conetadas. Se ligar uma região cortical aos olhos, obtém visão; se ligar a mesma região aos ouvidos, obtém audição; se ligar regiões a outras regiões, obtém pensamentos mais elevados, como a linguagem. Mountcastle conclui que, se conseguirmos descobrir a função básica de qualquer parte do neocórtex, entenderemos como o todo funciona.

A ideia de Mountcastle é tão surpreendente e profunda quanto a descoberta da evolução por Darwin. Darwin propôs um mecanismo — um algoritmo, por assim dizer — que explica a incrível diversidade da vida. O que à superfície parece ser uma multiplicidade de animais e plantas, muitos tipos diferentes de seres vivos, são, na realidade, manifestações do mesmo algoritmo

evolutivo subjacente. Por sua vez, Mountcastle propõe que todas as coisas que associamos à inteligência, que à superfície parecem diferentes, são, na realidade, manifestações do mesmo algoritmo cortical subjacente. Espero que possa perceber o quão inesperada e revolucionária é a proposta de Mountcastle. Darwin propôs que a diversidade da vida resulta de um algoritmo básico. Mountcastle propôs que a diversidade da inteligência também resulta de um algoritmo básico.

Como acontece com muitas ideias historicamente significativas, há algum debate sobre se Mountcastle foi a primeira pessoa a propor esta ideia. A minha experiência é que toda a ideia tem pelo menos algum precedente. Mas, tanto quanto sei, Mountcastle foi o primeiro a expor clara e cuidadosamente o argumento de um algoritmo cortical comum.

As propostas de Mountcastle e Darwin diferem num aspeto interessante. Darwin sabia qual era o algoritmo: a evolução baseia-se na variação aleatória e na seleção natural. No entanto, Darwin desconhecia onde estava esse algoritmo no corpo — algo que só foi descoberto muitos anos depois, com a descoberta do DNA. Mountcastle, por outro lado, não sabia qual era o algoritmo cortical; não conhecia os princípios da inteligência. Mas sabia onde esse algoritmo residia no cérebro.

Então, qual foi a proposta de Mountcastle para a localização do algoritmo cortical? Ele disse que a unidade fundamental do neocórtex, a unidade da inteligência, era a "coluna cortical". Observando a superfície do neocórtex, uma coluna cortical ocupa cerca de um milímetro quadrado. Estende-se por toda a espessura de 2,5 mm, o que lhe confere um volume de 2,5 milímetros cúbicos. Por esta definição, existem aproximadamente 150 000 colunas corticais empilhadas lado a lado num neocórtex humano. Pode imaginar uma coluna cortical como um pequeno pedaço de esparguete fino. Um neocórtex humano é como 150 000 pedaços curtos de esparguete empilhados verticalmente lado a lado.

A largura das colunas corticais varia entre espécies e regiões. Por exemplo, em ratos e camundongos, há uma coluna cortical para cada bigode; estas colunas têm cerca de meio milímetro de diâmetro. Nos gatos, as colunas visuais parecem ter cerca de um

milímetro de diâmetro. Não temos muitos dados sobre o tamanho das colunas no cérebro humano. Para simplificar, continuarei a referir-me às colunas como tendo um milímetro quadrado, dotando cada um de nós de cerca de 150 000 colunas corticais. Mesmo que o número real possa variar, isso não fará diferença para os nossos propósitos.

As colunas corticais não são visíveis ao microscópio. Com algumas exceções, não existem limites visíveis entre elas. Os cientistas sabem que existem porque todas as células numa coluna respondem à mesma parte da retina ou ao mesmo pedaço de pele, enquanto as células na coluna adjacente respondem a uma parte diferente da retina ou a outra área da pele. Este agrupamento de respostas é o que define uma coluna. Esta organização está presente em todo o neocórtex. Mountcastle salientou que cada coluna é ainda dividida em várias centenas de "minicolunas". Se uma coluna cortical for como um fio fino de esparguete, pode imaginar as minicolunas como fios ainda mais finos, semelhantes a cabelos individuais, empilhados, lado a lado, dentro do fio de esparguete. Cada minicoluna contém um pouco mais de cem neurónios, abrangendo todas as camadas. Ao contrário da maior coluna cortical, as minicolunas são fisicamente distintas e frequentemente visíveis ao microscópio.

Mountcastle não sabia nem sugeriu qual seria a função das colunas ou minicolunas. Propôs apenas que cada coluna realiza a mesma função e que as minicolunas são um subcomponente importante.

Vamos recapitular. O neocórtex é uma folha de tecido do tamanho de um grande guardanapo. Está dividido em dezenas de regiões que desempenham funções diferentes. Cada região divide-se em milhares de colunas. Cada coluna é composta por várias centenas de minicolunas, semelhantes a cabelos, que por sua vez contêm pouco mais de cem células cada. Mountcastle propôs que, em todo o neocórtex, colunas e minicolunas realizam a mesma função: implementar um algoritmo fundamental responsável por todos os aspetos da perceção e da inteligência.

Mountcastle baseou a sua proposta de um algoritmo universal em várias evidências. Primeiro, como já referi, os circuitos

detalhados do neocórtex são surpreendentemente semelhantes em todas as regiões. Se eu lhe mostrasse dois chips de silício com circuitos quase idênticos, seria seguro assumir que desempenham funções quase idênticas. O mesmo raciocínio aplica-se aos circuitos detalhados do neocórtex. Em segundo lugar, a grande expansão do neocórtex humano moderno em relação aos nossos antepassados hominídeos ocorreu rapidamente, numa escala evolutiva de apenas alguns milhões de anos. Este é provavelmente um período demasiado curto para a evolução desenvolver múltiplas novas capacidades complexas, mas é tempo suficiente para fazer mais cópias do mesmo circuito básico. Em terceiro lugar, a função das regiões neocorticais não está fixa. Por exemplo, em pessoas com cegueira congénita, as áreas visuais do neocórtex não recebem informação útil dos olhos, pelo que essas áreas podem assumir novas funções relacionadas com a audição ou o tato. Finalmente, há o argumento da extrema flexibilidade. Os humanos conseguem realizar muitas coisas para as quais não existiu pressão evolutiva, como programar computadores ou fazer gelados — invenções recentes. O facto de podermos fazer estas coisas indica que o cérebro depende de um método de aprendizagem de propósito geral. Para mim, este último argumento é o mais convincente: a capacidade de aprender praticamente tudo exige que o cérebro opere segundo um princípio universal.

Existem mais evidências que apoiam a proposta de Mountcastle. No entanto, apesar disso, a sua ideia foi controversa quando a apresentou e ainda hoje permanece algo controversa. Acredito que existem duas razões relacionadas para isso. Uma é que Mountcastle não sabia o que exatamente faz uma coluna cortical. Fez uma afirmação surpreendente baseada em muitas evidências circunstanciais, mas não propôs como uma coluna cortical poderia, na prática, realizar todas as funções que associamos à inteligência. A outra razão é que as implicações da sua proposta são difíceis de aceitar para algumas pessoas. Por exemplo, pode ser difícil aceitar que a visão e a linguagem sejam fundamentalmente a mesma coisa, pois elas não parecem semelhantes na experiência. Dadas estas incertezas, alguns cientistas rejeitam a proposta de Mountcastle apontando as diferenças entre as regiões do neocórtex. Embora essas diferenças

sejam relativamente pequenas comparadas com as semelhanças, quem as destaca pode argumentar que diferentes regiões do neocórtex não são iguais.

A proposta de Mountcastle paira na neurociência como um santo graal. Independentemente do animal ou da região cerebral que um neurocientista estude, quase todos, aberta ou discretamente, desejam entender como funciona o cérebro humano. E isso significa compreender como funciona o neocórtex. E isso requer entender o que faz uma coluna cortical. No fim, a nossa busca para compreender o cérebro, a nossa busca para entender a inteligência, resume-se a descobrir o que uma coluna cortical faz e como o faz. As colunas corticais não são o único mistério do cérebro nem o único mistério relacionado ao neocórtex, mas compreender a coluna cortical é, de longe, a maior e mais importante peça deste quebra-cabeças.

Em 2005, fui convidado a dar uma palestra sobre a nossa investigação na Universidade Johns Hopkins. Falei sobre a nossa busca por compreender o neocórtex, sobre a forma como estávamos a abordar o problema e os progressos que já tínhamos feito. Após uma palestra deste tipo, é comum o orador reunir-se com membros individuais do corpo docente. Nesta visita, o meu último encontro foi com Vernon Mountcastle e com o diretor do seu departamento. Senti-me honrado por conhecer o homem que me tinha proporcionado tanta visão e inspiração ao longo da vida. A certa altura da nossa conversa, Mountcastle, que assistira à minha palestra, disse que eu devia ir trabalhar para Johns Hopkins e que se encarregaria de me arranjar uma posição. A sua proposta foi inesperada e pouco habitual. Não a podia considerar seriamente devido aos meus compromissos familiares e profissionais na Califórnia, mas lembrei-me de 1986, quando a minha proposta para estudar o neocórtex foi rejeitada pela UC Berkeley. Como eu teria aceitado de imediato a oferta dele, naquela altura!

Antes de partir, pedi a Mountcastle que autografasse o meu exemplar já muito lido de *The Mindful Brain*. Ao afastar-me, sentia-me ao mesmo tempo feliz e triste. Feliz por o ter conhecido e aliviado por ele pensar bem de mim. Triste por saber que era possível nunca mais o voltar a ver. Mesmo que eu viesse a ter êxito na minha busca, talvez já não pudesse partilhar com ele aquilo que tinha aprendido, nem obter a sua ajuda e opinião. Enquanto caminhava para o táxi, senti-me determinado a completar a sua missão.

CAPÍTULO 3

Um Modelo do Mundo na Sua Cabeça

O que o cérebro faz pode parecer-lhe óbvio. O cérebro recebe estímulos dos seus sensores, processa esses estímulos e depois age. No final de contas, a forma como um animal reage ao que sente determina o seu sucesso ou fracasso. Um mapeamento direto entre entrada sensorial e ação aplica-se certamente a algumas partes do cérebro. Por exemplo, tocar acidentalmente numa superfície quente provoca uma retração reflexa do braço. O circuito responsável por essa resposta está localizado na espinal medula. Mas e o neocórtex? Podemos afirmar que a função do neocórtex é receber estímulos dos sentidos e agir de imediato? Em suma, não.

Está a ler ou a ouvir este livro, e isso não está a provocar nenhuma ação imediata, para além de, talvez, virar páginas ou tocar num ecrã. Milhares de palavras estão a fluir para o seu neocórtex e, na maior parte das vezes, não está a agir em função delas. Talvez mais tarde atue de forma diferente por ter lido este livro. Talvez venha a ter conversas futuras sobre teoria do cérebro e o futuro da humanidade que não teria tido se não o tivesse lido. Talvez os seus pensamentos e escolhas de palavras, no futuro, sejam subtilmente influenciados pelas minhas palavras. Talvez até venha a trabalhar na criação de máquinas inteligentes baseadas em princípios do funcionamento cerebral, e as minhas palavras o inspirem nesse sentido. Mas, neste momento, está apenas a ler. Se insistirmos em descrever o neocórtex como um sistema de entrada e saída, então o melhor que podemos dizer é que o neocórtex recebe inúmeros estímulos, aprende com eles e, mais tarde — talvez horas, talvez anos depois — age de forma diferente com base nesses estímulos anteriores.

Desde o momento em que me interessei por compreender o funcionamento do cérebro, percebi que pensar no neocórtex como

um sistema do tipo "entrada leva a saída" não seria frutífero. Felizmente, enquanto era estudante de doutoramento em Berkeley, tive uma intuição que me conduziu por um caminho diferente e mais bem-sucedido. Estava em casa, a trabalhar à secretária. Havia dezenas de objetos sobre a secretária e na sala. E apercebi-me de que, se qualquer um desses objetos mudasse, mesmo que de forma ínfima, eu notaria. O meu copo dos lápis estava sempre do lado direito da mesa; se um dia o encontrasse do lado esquerdo, notaria a mudança e questionar-me-ia sobre como foi lá parar. Se o agrafador alterasse ligeiramente o comprimento, eu repararia. Notaria essa mudança quer ao tocar no agrafador, quer ao olhá-lo. Notaria até se o som que faz ao ser utilizado fosse diferente. Se o relógio na parede mudasse de sítio ou de estilo, eu daria conta disso. Se o cursor no ecrã do computador se deslocasse para a esquerda quando movesse o rato para a direita, eu perceberia de imediato que algo estava errado. O que me impressionou foi o facto de eu reparar nestas mudanças mesmo quando não estava a prestar atenção a esses objetos. Ao olhar em redor da sala, eu não perguntava: "O meu agrafador tem o comprimento certo?" ou "Verifica se o ponteiro das horas continua mais curto do que o dos minutos". As alterações ao normal surgiam simplesmente na minha mente, e a minha atenção era então captada por elas. Literalmente milhares de mudanças possíveis no ambiente seriam notadas quase instantaneamente pelo meu cérebro.

Só havia uma explicação que me ocorria. O meu cérebro, especificamente o meu neocórtex, estava a fazer múltiplas previsões simultâneas sobre aquilo que estava prestes a ver, ouvir e sentir. Cada vez que eu movia os olhos, o meu neocórtex fazia previsões sobre o que estava prestes a ver. Cada vez que pegava em algo, fazia previsões sobre o que cada dedo deveria sentir. E cada ação que realizava levava a previsões sobre o que deveria ouvir. O meu cérebro previa desde os estímulos mais subtis — como a textura do cabo da minha chávena de café — até ideias conceptuais mais abrangentes — como o mês correto que deveria estar afixado num calendário. Essas previsões ocorriam em todas as modalidades sensoriais, tanto para características sensoriais de baixo nível como para conceitos de alto nível, o que me indicava que todas as partes do neocórtex, e, portanto, cada coluna cortical,

estavam a fazer previsões. A previsão era uma função ubíqua do neocórtex.

Naquela altura, poucos neurocientistas descreviam o cérebro como uma máquina de previsões. Focar-me em como o neocórtex realizava muitas previsões em paralelo parecia-me uma forma inovadora de estudar o seu funcionamento. Eu sabia que a previsão não era a única função do neocórtex, mas representava uma abordagem sistémica para enfrentar os mistérios da coluna cortical. Podia colocar perguntas específicas sobre como os neurónios fazem previsões em diferentes condições. As respostas a essas perguntas poderiam revelar o que fazem as colunas corticais — e como o fazem.

Para fazer previsões, o cérebro tem de aprender o que é "normal" — isto é, o que deve ser esperado com base na experiência passada. No meu livro anterior, *Sobre Inteligência*, explorei esta ideia da aprendizagem e da previsão. Nesse livro, utilizei a expressão "*estrutura de previsão baseada na memória*" para descrever o conceito geral e escrevi sobre as implicações de pensar o cérebro dessa forma. Defendi que, ao estudar como o neocórtex faz previsões, seríamos capazes de desvendar o seu funcionamento.

Hoje, já não utilizo a expressão "*estrutura de previsão baseada na memória*". Em vez disso, descrevo a mesma ideia dizendo que o neocórtex aprende um modelo do mundo, e faz previsões com base nesse modelo. Prefiro a palavra "modelo" porque descreve de forma mais precisa o tipo de informação que o neocórtex aprende. Por exemplo, o meu cérebro tem um modelo do meu agrafador. Esse modelo inclui o seu aspeto visual, a sensação ao toque e os sons que produz quando é usado. O modelo do mundo que o cérebro constrói inclui a localização dos objetos e a forma como eles mudam quando interagimos com eles. Por exemplo, o meu modelo do agrafador inclui a forma como a parte superior se move em relação à inferior e como o agrafo sai quando pressionamos o topo. Estas ações podem parecer simples, mas não nascemos a saber isto. Aprendemo-lo algures ao longo da vida, e agora essa informação está armazenada no nosso neocórtex.

O cérebro cria um *modelo preditivo*. Isto significa apenas que o cérebro está continuamente a prever quais serão os seus estímulos. A previsão não é algo que o cérebro faça de vez em quando; é uma propriedade intrínseca que nunca cessa, e cumpre um papel essencial na aprendizagem. Quando as previsões do cérebro se confirmam, isso significa que o seu modelo do mundo é preciso. Um erro de previsão chama a atenção para a discrepância e conduz à atualização do modelo.

Não temos consciência da esmagadora maioria destas previsões, a menos que o estímulo recebido pelo cérebro não corresponda ao esperado. Quando estendo casualmente a mão para pegar na minha chávena de café, não tenho noção de que o meu cérebro está a prever o que cada dedo deverá sentir, qual deverá ser o peso da chávena, a sua temperatura, e o som que deverá emitir quando a pousar novamente na secretária. Mas, se a chávena estivesse subitamente mais pesada, ou fria, ou emitisse um rangido, eu notaria a diferença. Podemos ter a certeza de que essas previsões estão a ocorrer porque mesmo uma pequena alteração em qualquer desses estímulos é notada. Porém, quando a previsão está correta — como acontece na maioria das vezes — não temos qualquer consciência de que ela ocorreu.

Ao nascermos, o nosso neocórtex praticamente nada sabe. Não conhece palavras, não sabe como são os edifícios, como se utiliza um computador, nem o que é uma porta ou como ela se move nas dobradiças. Tem de aprender uma infinidade de coisas. A estrutura geral do neocórtex não é aleatória. O seu tamanho, o número de regiões que o compõem e a forma como estão interligadas são, em grande parte, determinados pelos nossos genes. Por exemplo, os genes determinam que partes do neocórtex estão ligadas aos olhos, que outras estão ligadas aos ouvidos, e como essas zonas se conetam entre si. Podemos, assim, afirmar que o neocórtex nasce estruturado para ver, ouvir e até aprender linguagem. Mas também é verdade que não sabe, à partida, o que irá ver, o que irá ouvir, nem que línguas concretas poderá vir a aprender. Podemos imaginar o neocórtex como iniciando a vida com alguns pressupostos inatos sobre o mundo, mas sem conhecimento específico. Ao longo da experiência, vai construindo um modelo do mundo rico e complexo.

A quantidade de coisas que o neocórtex aprende é imensa. Estou sentado numa sala com centenas de objetos. Vou escolher um ao acaso: uma impressora. Aprendi um modelo da impressora que inclui a existência de uma bandeja de papel, e como essa bandeja se move para dentro e para fora do aparelho. Sei como alterar o tamanho do papel, como desembrulhar uma resma nova e colocá-la na bandeja. Sei os passos que devo seguir para resolver um encravamento de papel. Sei que o cabo de alimentação tem uma ficha em forma de D numa das extremidades e que só pode ser inserida numa determinada orientação. Reconheço o som da impressora e sei distingui-lo quando imprime em frente e verso ou apenas numa face da folha. Outro objeto na sala é um pequeno armário de arquivo com duas gavetas. Consigo recordar-me de dezenas de coisas que sei sobre esse armário, incluindo o conteúdo de cada gaveta e a forma como os objetos estão dispostos. Sei que tem uma fechadura, onde está a chave, e como introduzir e rodar a chave para trancar o armário. Sei como são a sensação e os sons ao usar a chave e a fechadura. A chave tem um pequeno anel metálico, e sei como usar a unha para abrir esse anel e adicionar ou remover chaves.

Imagine que percorre, de divisão em divisão, a sua casa. Em cada uma delas, consegue evocar centenas de elementos e, para cada item, pode seguir uma cascata de conhecimentos adquiridos. Pode fazer o mesmo exercício com a cidade onde vive, recordando que edifícios, parques, suportes para bicicletas e árvores específicas existem em diferentes locais. Para cada elemento, consegue lembrar-se de experiências associadas e de como interage com ele. A quantidade de coisas que sabe é enorme, e os elos de conhecimento associados parecem não ter fim. Aprendemos também muitos conceitos de alto nível. Estima-se que cada um de nós conheça cerca de quarenta mil palavras. Temos a capacidade de aprender linguagem falada, linguagem escrita, linguagem gestual, a linguagem da matemática e a linguagem da música. Aprendemos como funcionam formulários eletrónicos, o que fazem os termóstatos, e até o que significam conceitos como empatia ou democracia — ainda que a nossa compreensão destes possa variar. Independentemente de outras funções que o neocórtex possa desempenhar, podemos afirmar com certeza que ele aprende um

modelo incrivelmente complexo do mundo. Este modelo é a base das nossas previsões, perceções e ações.

Aprendizagem Através do Movimento

Os estímulos que chegam ao cérebro estão em constante mudança. Há duas razões para isso. Primeiro, o mundo pode mudar. Por exemplo, ao ouvir música, os estímulos provenientes dos ouvidos alteram-se rapidamente, acompanhando o movimento da melodia. De forma semelhante, uma árvore que oscila com a brisa provocará alterações visuais e, talvez, auditivas. Nestes dois exemplos, os estímulos dirigidos ao cérebro mudam de momento para momento, não porque estejamos em movimento, mas porque são as coisas no mundo que se estão a mover ou a transformar por si mesmas.

A segunda razão é que nós próprios nos movemos. Cada vez que damos um passo, movemos um membro, desviamos o olhar, inclinamos a cabeça ou emitimos um som, os estímulos sensoriais que recebemos mudam. Por exemplo, os nossos olhos fazem movimentos rápidos, chamados sacádicos, cerca de três vezes por segundo. Com cada sacada, os olhos fixam-se num novo ponto do mundo, e a informação visual enviada ao cérebro muda completamente. Essa alteração não ocorreria se não tivéssemos movido os olhos.

O cérebro aprende o seu modelo do mundo observando como os seus estímulos variam ao longo do tempo. Não há outra forma de aprender. Ao contrário de um computador, não podemos carregar um ficheiro para dentro do nosso cérebro. A única maneira de o cérebro aprender seja o que for é através das mudanças nos seus estímulos. Se os estímulos fossem estáticos, nada poderia ser aprendido.

Algumas coisas, como uma melodia, podem ser aprendidas sem movimento corporal. Podemos ficar perfeitamente imóveis, com os olhos fechados, e aprender uma nova melodia apenas ouvindo como os sons mudam com o tempo. Mas a maior parte da aprendizagem requer que nos movamos ativamente e exploremos. Imagine que entra numa casa nova, onde nunca esteve antes. Se

não se mexer, não haverá alterações nos seus estímulos sensoriais, e não poderá aprender absolutamente nada sobre a casa. Para construir um modelo da casa, tem de olhar em direções diferentes e percorrer as várias divisões. Tem de abrir portas, espreitar gavetas e pegar em objetos. A casa e o seu conteúdo são, na maioria dos casos, estáticos — não se movem por iniciativa própria. Para aprender um modelo de uma casa, é necessário movimento.

Peguemos num objeto simples, como um rato de computador. Para aprender como é o tato do rato, tem de passar os dedos pela sua superfície. Para aprender como é o seu aspeto, tem de observá-lo de vários ângulos e fixar os olhos em diferentes pontos. Para compreender o que faz, tem de pressionar os botões, deslizar a tampa das pilhas ou movê-lo sobre um tapete, observando, sentindo e ouvindo o que acontece.

O termo que descreve isto é aprendizagem sensório-motora. Por outras palavras, o cérebro constrói um modelo do mundo observando como os estímulos sensoriais mudam à medida que nos movemos. Podemos aprender uma canção sem nos movermos porque, ao contrário da ordem com que percorremos uma casa, a ordem das notas numa melodia é fixa. Mas a maior parte do mundo não é assim; na maioria das vezes temos de nos mover para descobrir a estrutura dos objetos, dos lugares e das ações. Na aprendizagem sensório-motora, ao contrário de uma melodia, a sequência das sensações não é fixa. O que vejo ao entrar numa divisão depende da direção para a qual viro a cabeça. O que o meu dedo sente ao segurar numa chávena de café depende de se o mover para cima, para baixo ou de lado.

A cada movimento, o neocórtex prevê qual será a próxima sensação. Se mover o meu dedo para cima, ao longo da chávena de café, espero sentir a borda; se o mover para o lado, espero tocar na asa. Se virar a cabeça para a esquerda ao entrar na cozinha, espero ver o frigorífico; se a virar para a direita, espero ver o fogão. Se direcionar o olhar para o queimador da frente, do lado esquerdo, espero ver o acendedor partido que preciso de consertar. Se algum estímulo não corresponder à previsão do cérebro — por exemplo, se o meu cônjuge tiver arranjado o acendedor — então a minha atenção será automaticamente dirigida para essa discrepância. Isso

alerta o neocórtex de que o seu modelo daquela parte do mundo precisa de ser atualizado.

A questão sobre como funciona o neocórtex pode agora ser formulada de forma mais precisa: Como é que o neocórtex, composto por milhares de colunas corticais quase idênticas, aprende um modelo preditivo do mundo através do movimento?

Foi esta a pergunta que a minha equipa e eu nos propusemos responder. Acreditávamos que, se conseguíssemos dar resposta a esta questão, poderíamos fazer engenharia reversa do neocórtex. Compreenderíamos tanto o que o neocórtex faz como, como o faz. E, em última instância, seríamos capazes de construir máquinas que funcionassem da mesma maneira.

Dois Princípios da Neurociência

Antes de podermos começar a responder à questão acima, há ainda algumas ideias fundamentais que precisa de conhecer. Em primeiro lugar, tal como todas as outras partes do corpo, o cérebro é composto por células. As células do cérebro, chamadas neurónios, são em muitos aspetos semelhantes às restantes células do nosso organismo. Por exemplo, um neurónio possui uma membrana celular que define os seus limites, e um núcleo que contém ADN. No entanto, os neurónios apresentam várias propriedades únicas que não existem noutras células do corpo.

A primeira é que os neurónios têm uma aparência semelhante à de árvores. Possuem extensões ramificadas da membrana celular, chamadas axónios e dendrites. As ramificações das dendrites estão agrupadas em torno da célula e recolhem os estímulos recebidos. O axónio é a via de saída: transmite os sinais e estabelece múltiplas ligações com neurónios vizinhos, podendo por vezes percorrer longas distâncias — por exemplo, de um lado ao outro do cérebro ou do neocórtex até à espinal medula.

A segunda diferença é que os neurónios produzem picos elétricos, também chamados potenciais de ação. Um potencial de ação é um sinal elétrico que se inicia junto ao corpo celular e percorre o axónio até atingir a extremidade de cada uma das suas ramificações.

A terceira propriedade única é que o axónio de um neurónio estabelece ligações com as dendrites de outros neurónios. Os pontos de ligação são chamados sinapses. Quando um pico elétrico percorre o axónio e chega a uma sinapse, liberta uma substância química que entra na dendrite do neurónio receptor. Dependendo do tipo de substância libertada, essa ação torna o neurónio receptor mais ou menos propenso a gerar o seu próprio pico elétrico.

Tendo em conta o funcionamento dos neurónios, podemos enunciar dois princípios fundamentais. Estes princípios desempenharão papéis essenciais na nossa compreensão do cérebro e da inteligência.

Princípio Número Um: Pensamentos, Ideias e Perceções São a Atividade dos Neurónios

A qualquer momento, alguns neurónios no neocórtex estão ativamente a disparar e outros não. Tipicamente, a percentagem de neurónios ativos em simultâneo é reduzida — talvez apenas 2 por cento. Os seus pensamentos e perceções são determinados por quais neurónios estão a disparar. Por exemplo, durante uma cirurgia cerebral, os médicos, por vezes, precisam de ativar neurónios no cérebro de um paciente acordado. Introduzem uma sonda minúscula no neocórtex e utilizam eletricidade para ativar um pequeno grupo de neurónios. Quando o fazem, o paciente pode ouvir, ver ou pensar em algo. Quando a estimulação cessa, a experiência que o paciente estava a ter termina também. Se o médico ativar um conjunto diferente de neurónios, o paciente terá um pensamento ou perceção distintos.

Os pensamentos e as experiências são sempre o resultado de um determinado conjunto de neurónios ativos em simultâneo. Um mesmo neurónio pode participar em muitos pensamentos ou experiências diferentes. Cada pensamento que tem é a atividade de neurónios. Tudo o que vê, ouve ou sente é igualmente atividade neuronal. Os nossos estados mentais e a atividade dos neurónios são, de facto, uma e a mesma coisa.

Princípio Número Dois: Tudo o que Sabemos Está Armazenado nas Ligações Entre os Neurónios

O cérebro retém uma enorme quantidade de informação. Tem memórias permanentes, como o local onde cresceu. Tem memórias temporárias, como o que jantou ontem. E possui conhecimentos básicos, como a forma de abrir uma porta ou como se soletra a palavra "dicionário". Todas estas coisas são armazenadas através das sinapses — as ligações entre neurónios.

Eis a ideia fundamental de como o cérebro aprende: cada neurónio possui milhares de sinapses, que o ligam a milhares de outros neurónios. Se dois neurónios disparam ao mesmo tempo, a ligação entre eles é reforçada. Quando aprendemos algo, essas conexões tornam-se mais fortes; quando esquecemos, enfraquecem. Esta ideia básica foi proposta por Donald Hebb nos anos 1940, e hoje é conhecida como *aprendizagem hebbiana*.

Durante muitos anos, acreditou-se que as ligações entre neurónios num cérebro adulto eram fixas. Supunha-se que aprender consistia apenas em aumentar ou diminuir a força das sinapses. Esta é, ainda hoje, a forma como ocorre a aprendizagem na maioria das redes neurais artificiais.

Contudo, nas últimas décadas, os cientistas descobriram que, em muitas zonas do cérebro — incluindo o neocórtex — novas sinapses se formam e outras desaparecem. Todos os dias, muitas das sinapses de um neurónio são eliminadas e substituídas por novas. Assim, grande parte da aprendizagem acontece através da formação de novas conexões entre neurónios que antes não estavam ligados. O esquecimento dá-se quando essas ligações antigas ou não utilizadas são removidas por completo.

As conexões no nosso cérebro armazenam o modelo do mundo que aprendemos através da experiência. Todos os dias vivenciamos novas situações e acrescentamos novos fragmentos de conhecimento ao modelo, criando novas sinapses. Os neurónios ativos num dado momento representam os nossos pensamentos e perceções atuais.

Chegados aqui, percorremos já vários dos blocos fundamentais que compõem o neocórtex — algumas das peças do nosso quebra-

cabeças. No próximo capítulo, começaremos a encaixá-las para revelar como funciona, no seu todo, o neocórtex.

CAPÍTULO 4

O Cérebro Revela os Seus Segredos

As pessoas costumam dizer que o cérebro é a coisa mais complexa do universo. E concluem, a partir disso, que não haverá uma explicação simples para o modo como funciona, ou que talvez nunca o venhamos a compreender. A história da descoberta científica sugere que estão enganadas. As grandes descobertas são quase sempre precedidas por observações desconcertantes e complexas. Com o enquadramento teórico correto, a complexidade não desaparece, mas deixa de parecer confusa ou intimidante.

Um exemplo familiar é o movimento dos planetas. Durante milhares de anos, os astrónomos acompanharam cuidadosamente o movimento dos planetas entre as estrelas. O percurso de um planeta ao longo de um ano é complexo, ziguezagueando para cá e para lá, traçando laços no céu. Era difícil imaginar uma explicação para esses movimentos desordenados. Hoje, todas as crianças aprendem a ideia básica de que os planetas orbitam o Sol. O movimento dos planetas continua a ser complexo, e prever o seu curso exige matemática difícil, mas com o enquadramento certo, a complexidade já não é misteriosa. Poucas descobertas científicas são difíceis de compreender num nível elementar. Uma criança pode aprender que a Terra orbita o Sol. Um aluno do ensino secundário pode aprender os princípios da evolução, da genética, da mecânica quântica e da relatividade. Cada um destes avanços científicos foi precedido por observações confusas. Mas hoje parecem claros e lógicos.

Do mesmo modo, sempre acreditei que o neocórtex parecia complicado sobretudo porque não o compreendíamos, e que pareceria relativamente simples em retrospetiva. Uma vez descoberta a solução, olharíamos para trás e diríamos: “Ah, claro, como é que não pensámos nisso antes?” Quando a nossa investigação estagnava ou quando me diziam que o cérebro era

demasiado difícil de entender, eu imaginava um futuro em que a teoria do cérebro faria parte dos programas escolares do ensino secundário. Isso mantinha-me motivado.

O nosso progresso na tentativa de decifrar o neocórtex teve altos e baixos. Ao longo de dezoito anos — três no Instituto de Neurociência de Redwood e quinze na Numenta — os meus colegas e eu trabalhámos nesse problema. Houve alturas em que fizemos pequenos avanços, outras em que obtivemos grandes avanços, e outras ainda em que seguimos ideias que, a princípio, pareciam promissoras, mas que, no fim, se revelaram becos sem saída. Não lhe vou relatar toda essa história. Em vez disso, quero descrever vários momentos-chave em que a nossa compreensão deu um salto, em que a natureza sussurrou ao nosso ouvido algo que tínhamos negligenciado. Há três desses momentos de "eureka" que recordo vividamente.

Descoberta Número Um: O Neocórtex Aprende um Modelo Preditivo do Mundo

Já descrevi como, em 1986, percebi que o neocórtex aprende um modelo preditivo do mundo. Não consigo exagerar a importância desta ideia. Chamo-lhe uma descoberta porque foi assim que a senti na altura. Há uma longa história de filósofos e cientistas a falar de ideias relacionadas, e hoje não é incomum que neurocientistas digam que o cérebro aprende um modelo preditivo do mundo. Mas, em 1986, os neurocientistas e os manuais escolares ainda descreviam o cérebro de forma mais semelhante a um computador: a informação entra, é processada, e depois o cérebro atua. Claro que aprender um modelo do mundo e fazer previsões não é a única coisa que o neocórtex faz. No entanto, ao estudar como o neocórtex faz previsões, acreditava que poderíamos desvendar o funcionamento de todo o sistema.

Esta descoberta levou a uma pergunta importante: como é que o cérebro faz previsões? Uma possível resposta é que o cérebro tem dois tipos de neurónios: neurónios que disparam quando o cérebro está efetivamente a ver algo, e neurónios que disparam quando o cérebro está a prever que verá algo. Para evitar alucinações, o

cérebro precisa de manter as suas previsões separadas da realidade. Utilizar dois conjuntos de neurónios cumpre bem essa função. No entanto, há dois problemas com esta ideia.

Primeiro, dado que o neocórtex está a fazer um número massivo de previsões a cada momento, esperaríamos encontrar um grande número de neurónios preditivos. Até agora, isso não foi observado. Os cientistas encontraram alguns neurónios que se tornam ativos antes de um estímulo, mas estes neurónios não são tão comuns como esperaríamos. O segundo problema baseia-se numa observação que há muito me incomodava. Se o neocórtex está a fazer centenas ou milhares de previsões a cada momento, por que não estamos conscientes da maioria dessas previsões? Se pego numa chávena com a mão, não estou consciente de que o meu cérebro está a prever o que cada dedo deverá sentir, a não ser que sinta algo invulgar — por exemplo, uma fissura. Não temos consciência da maioria das previsões feitas pelo cérebro, a não ser quando ocorre um erro. A tentativa de compreender como os neurónios no neocórtex fazem previsões conduziu à segunda descoberta.

Descoberta Número Dois: As Previsões Ocorrem no Interior dos Neurónios

Recordemos que as previsões feitas pelo neocórtex ocorrem sob duas formas. Um tipo ocorre porque o mundo está a mudar à sua volta. Por exemplo, está a ouvir uma melodia. Pode estar sentado, imóvel, de olhos fechados, e o som que entra pelos seus ouvidos muda à medida que a melodia avança. Se conhecer a melodia, o seu cérebro antecipa continuamente a nota seguinte, e dará conta se alguma das notas estiver errada. O segundo tipo de previsão ocorre porque está a mover-se em relação ao mundo. Por exemplo, quando prendo a minha bicicleta no átrio do meu escritório, o meu neocórtex faz muitas previsões sobre o que irei sentir, ver e ouvir com base nos meus movimentos. A bicicleta e o cadeado não se movem por si mesmos. Cada ação que faço gera um conjunto de previsões. Se eu alterar a ordem das minhas ações, a ordem das previsões também muda.

A proposta de Mountcastle de um algoritmo cortical comum sugeria que cada coluna do neocórtex faz ambos os tipos de previsões. Caso contrário, as colunas corticais teriam funções diferentes. A minha equipa também percebeu que os dois tipos de previsão estão intimamente relacionados. Por isso, sentimos que o progresso num dos subproblemas levaria ao progresso no outro.

Prever a nota seguinte numa melodia — também conhecido como memória sequencial — é o mais simples dos dois problemas, por isso começámos por aí. A memória sequencial é usada para muito mais do que aprender melodias; é também usada na criação de comportamentos. Por exemplo, quando me seco com uma toalha depois de tomar banho, sigo tipicamente um padrão de movimentos quase idêntico, o que é uma forma de memória sequencial. A memória sequencial também é usada na linguagem. Reconhecer uma palavra falada é como reconhecer uma melodia curta. A palavra é definida por uma sequência de fonemas, enquanto a melodia é definida por uma sequência de intervalos musicais. Há muitos mais exemplos, mas para simplificar, vou cingir-me às melodias. Ao deduzir como os neurónios numa coluna cortical aprendem sequências, esperávamos descobrir princípios básicos sobre como os neurónios fazem previsões sobre tudo.

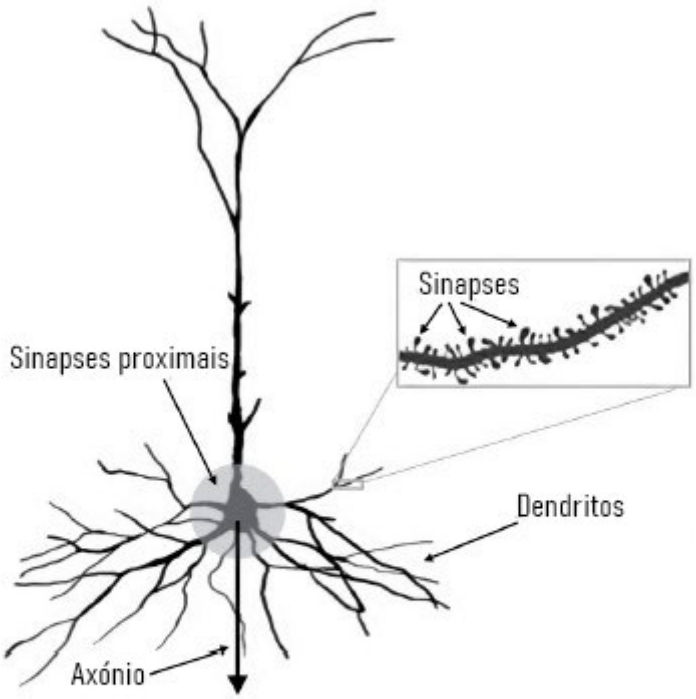
Trabalhámos no problema da previsão de melodias durante vários anos antes de conseguirmos deduzir a solução, a qual tinha de apresentar inúmeras capacidades. Por exemplo, as melodias têm frequentemente secções repetidas, como um refrão ou o *da da da dum* da Quinta Sinfonia de Beethoven. Para prever a nota seguinte, não se pode olhar apenas para a nota anterior ou para as últimas cinco notas. A previsão correta pode depender de notas que ocorreram há muito tempo. Os neurónios têm de perceber quanta contextualização é necessária para fazer a previsão certa. Outro requisito é que os neurónios têm de jogar ao "Adivinha a Canção". As primeiras notas que ouve podem pertencer a várias melodias diferentes. Os neurónios têm de acompanhar todas as melodias possíveis que sejam compatíveis com o que foi ouvido até ao momento, até que notas suficientes tenham sido ouvidas para eliminar todas, exceto uma.

Engenheirar uma solução para o problema da memória sequencial seria fácil, mas descobrir como os neurónios reais — organizados tal como os vemos no neocórtex — resolvem estes e outros requisitos foi difícil. Ao longo de vários anos, tentámos diferentes abordagens. A maioria funcionava até certo ponto, mas nenhuma apresentava todas as capacidades de que precisávamos e nenhuma se ajustava com precisão aos detalhes biológicos que conhecíamos sobre o cérebro. Não estávamos interessados numa solução parcial nem numa solução "inspirada biologicamente". Queríamos saber exatamente como é que os neurónios reais, organizados tal como são no neocórtex, aprendem sequências e fazem previsões.

Lembro-me do momento em que cheguei à solução do problema da previsão melódica. Foi em 2010, um dia antes do feriado do Dia de Ação de Graças. A solução surgiu num relâmpago. Mas, à medida que refletia sobre ela, percebi que exigia que os neurónios fizessem coisas que eu não tinha a certeza de que fossem capazes de fazer. Por outras palavras, a minha hipótese fazia várias previsões detalhadas e surpreendentes que eu podia testar.

Os cientistas normalmente testam uma teoria realizando experiências para ver se as previsões feitas pela teoria se confirmam ou não. Mas a neurociência é um campo invulgar. Existem centenas a milhares de artigos publicados em cada subárea, e a maioria desses artigos apresenta dados experimentais que não estão integrados em nenhuma teoria geral. Isso oferece a teóricos como eu a oportunidade de testar rapidamente uma nova hipótese pesquisando investigações anteriores em busca de dados experimentais que a sustentem ou refutem. Encontrei algumas dezenas de artigos científicos que continham dados experimentais capazes de lançar luz sobre a nova teoria da memória sequencial. A minha família alargada estava de visita para o feriado, mas eu estava demasiado entusiasmado para esperar até todos irem embora. Recordo-me de ler artigos enquanto cozinhava e de envolver os meus familiares em conversas sobre neurónios e melodias. Quanto mais lia, mais confiante ficava de que tinha descoberto algo importante.

A chave da descoberta foi uma nova forma de pensar os neurónios.



Um neurónio típico

Acima está uma imagem do tipo de neurónio mais comum no neocórtex. Neurónios como este têm milhares, por vezes dezenas de milhares, de sinapses distribuídas ao longo dos ramos dos dendritos. Alguns dos dendritos situam-se perto do corpo celular (que está mais próximo da parte inferior da imagem), e outros encontram-se mais afastados (mais próximo do topo). A caixa mostra uma vista ampliada de um ramo dendrítico, para que se possa ver quão pequenas e densamente agrupadas são as sinapses. Cada saliência ao longo do dendrito é uma sinapse. Realcei também uma zona à volta do corpo celular; as sinapses nesta área são chamadas sinapses proximais. Se as sinapses proximais recebem estímulo suficiente, o neurónio dispara. O disparo começa no corpo celular e viaja até outros neurónios através do axónio. O axónio não era visível nesta imagem, por isso adicionei uma seta voltada para baixo para indicar onde ele estaria. Se considerarmos apenas as sinapses proximais e o corpo celular, temos a visão clássica de um

neurónio. Se alguma vez leu sobre neurónios ou estudou redes neurais artificiais, reconhecerá esta descrição.

Curiosamente, menos de 10 por cento das sinapses da célula encontram-se na zona proximal. Os outros 90 por cento estão demasiado distantes para provocar um disparo. Se um estímulo chega a uma destas sinapses distais, como as que se veem na caixa, tem praticamente nenhum efeito no corpo celular. Tudo o que os investigadores conseguiam afirmar era que as sinapses distais desempenhavam algum tipo de papel modulador. Durante muitos anos, ninguém sabia ao certo para que serviam 90 por cento das sinapses do neocórtex.

A partir de cerca de 1990, esta visão começou a mudar. Os cientistas descobriram novos tipos de impulsos que se propagam ao longo dos dendritos. Antes, conhecíamos apenas um tipo de impulso: aquele que começava no corpo celular e viajava pelo axónio até alcançar outras células. Agora, descobrira-se que havia outros impulsos que percorriam os dendritos. Um tipo de impulso dendrítico começa quando um grupo de cerca de vinte sinapses situadas umas junto às outras num ramo dendrítico recebem estímulo ao mesmo tempo. Uma vez ativado o impulso dendrítico, este percorre o dendrito até alcançar o corpo celular. Quando lá chega, eleva a voltagem da célula, mas não o suficiente para fazer com que o neurónio dispare. É como se o impulso dendrítico estivesse a provocar o neurónio — é quase forte o suficiente para o ativar, mas não chega lá.

O neurónio permanece neste estado de provocação durante algum tempo antes de voltar ao normal. Os cientistas voltaram a ficar intrigados. Para que servem os impulsos dendríticos se não são suficientemente fortes para gerar um disparo no corpo celular? Como não se sabia para que serviam os impulsos dendríticos, os investigadores em inteligência artificial utilizam neurónios simulados que não os possuem. Esses neurónios também não têm dendritos, nem as milhares de sinapses que existem nos dendritos. Eu sabia que as sinapses distais tinham de desempenhar um papel essencial no funcionamento do cérebro. Qualquer teoria ou rede neural que não tivesse em conta 90 por cento das sinapses do cérebro tinha de estar errada.

A grande revelação que tive foi que os impulsos dendríticos são previsões. Um impulso dendrítico ocorre quando um conjunto de sinapses próximas umas das outras num dendrito distal recebem estímulo ao mesmo tempo, e isso significa que o neurónio reconheceu um padrão de atividade noutros neurónios. Quando esse padrão de atividade é detetado, gera-se um impulso dendrítico, que eleva a voltagem no corpo celular, colocando a célula no que chamamos *um estado preditivo*. O neurónio fica então preparado para disparar. É semelhante a um corredor que, ao ouvir 'Prontos, preparar...', se coloca nos blocos de partida, pronto a arrancar. Se um neurónio em estado preditivo recebe posteriormente estímulo suficiente nas sinapses proximais para gerar um potencial de ação, então dispara um pouco mais cedo do que teria disparado se não estivesse nesse estado preditivo.

Imagine que há dez neurónios que reconhecem todos o mesmo padrão nas suas sinapses proximais. É como se fossem dez corredores alinhados na linha de partida, todos à espera do mesmo sinal para começar a corrida. Um desses corredores ouve "Prontos, preparar..." e antecipa que a corrida está prestes a começar. Coloca-se nos blocos, pronto a arrancar. Quando ouve o sinal de partida, arranca antes dos outros que não estavam preparados, que não ouviram o sinal prévio. Ao verem o primeiro corredor já em vantagem, os outros desistem e nem sequer arrancam. Esperam pela próxima corrida. Este tipo de competição ocorre em todo o neocórtex.

Em cada minicoluna, múltiplos neurónios respondem ao mesmo padrão de estímulo. São como os corredores na linha de partida, todos à espera do mesmo sinal. Se o estímulo preferido chega, todos querem disparar. Contudo, se um ou mais desses neurónios estiverem em estado preditivo, a nossa teoria afirma que apenas esses neurónios disparam, enquanto os outros são inibidos. Assim, quando um estímulo inesperado chega, vários neurónios disparam em simultâneo. Se o estímulo for previsto, então apenas os neurónios em estado preditivo se tornam ativos. Esta é uma observação comum sobre o neocórtex: estímulos inesperados provocam muito mais atividade do que os esperados.

Se pegarmos em vários milhares de neurónios, os organizarmos em minicolunas, os deixarmos estabelecer ligações entre si, e adicionarmos alguns neurónios inibitórios, então esses neurónios resolvem o problema do "Advinha a Melodia", não se confundem com subsequências repetidas e, em conjunto, predizem o próximo elemento da sequência.

O segredo para que isto funcionasse foi uma nova compreensão do neurónio. Já sabíamos anteriormente que a previsão é uma função ubíqua do cérebro. Mas não sabíamos como, nem onde, as previsões eram feitas. Com esta descoberta, compreendemos que a maioria das previsões ocorre dentro dos próprios neurónios. Uma previsão acontece quando um neurónio reconhece um padrão, gera um impulso dendrítico, e fica preparado para disparar mais cedo do que outros neurónios. Com milhares de sinapses distais, cada neurónio pode reconhecer centenas de padrões que predizem quando deverá tornar-se ativo. A previsão está incorporada na própria estrutura do neocórtex — no neurónio.

Passámos mais de um ano a testar o novo modelo de neurónio e o circuito de memória de sequências. Escrevemos simulações em software para testar a sua capacidade e ficámos surpreendidos ao descobrir que, com apenas vinte mil neurónios, era possível aprender milhares de sequências completas. Descobrimos que a memória de sequência continuava a funcionar mesmo que 30 por cento dos neurónios morressem ou que o estímulo de entrada fosse ruidoso. Quanto mais tempo passávamos a testar a nossa teoria, mais confiança ganhávamos de que ela refletia realmente o que acontecia no neocórtex. Também encontrámos cada vez mais evidência empírica vinda de laboratórios experimentais que apoiava a nossa ideia. Por exemplo, a teoria previa que os impulsos dendríticos se comportariam de formas muito específicas, mas, inicialmente, não conseguíamos encontrar provas experimentais conclusivas. Contudo, ao falar com investigadores experimentais, conseguimos compreender melhor os seus resultados e perceber que os dados estavam de facto em consonância com as nossas previsões. Publicámos a teoria pela primeira vez num *white paper* em 2011. Depois, publicámos um artigo revisto por pares em 2016, intitulado "*Porque é que os Neurónios Têm Milhares de Sinapses: Uma Teoria da Memória Sequencial no Neocórtex*". A reação ao

artigo foi encorajadora, pois rapidamente se tornou o mais lido da sua revista científica.

Descoberta Número Três: O Segredo da Coluna Cortical São Referências

Em seguida, voltámos a nossa atenção para a segunda metade do problema da previsão: como é que o neocórtex prevê o próximo estímulo quando nos movemos? Ao contrário de uma melodia, a ordem dos estímulos nesta situação não é fixa, pois depende da direção do nosso movimento. Por exemplo, se olho para a esquerda, vejo uma coisa; se olho para a direita, vejo outra. Para que uma coluna cortical possa prever o seu próximo estímulo, tem de saber qual o movimento que está prestes a ocorrer.

Prever o próximo estímulo de uma sequência e prever o próximo estímulo ao movermo-nos são problemas semelhantes. Percebemos que o nosso circuito de memória sequencial poderia fazer ambos os tipos de previsão se os neurónios recebessem um estímulo adicional que representasse o movimento do sensor. No entanto, não sabíamos como esse sinal relacionado com o movimento deveria ser.

Começámos com a coisa mais simples que nos ocorreu: E se o sinal relacionado com o movimento fosse apenas "mover para a esquerda" ou "mover para a direita"? Testámos esta ideia — e funcionou. Chegámos até a construir um pequeno braço robótico que previa os seus estímulos à medida que se movia para a esquerda e para a direita, e apresentámo-lo numa conferência de neurociência. Contudo, o nosso braço robótico tinha limitações. Funcionava para problemas simples, como mover-se em duas direções, mas quando tentávamos escalá-lo para lidar com a complexidade do mundo real — como mover-se em múltiplas direções ao mesmo tempo — exigia treino em excesso. Sentíamos que estávamos próximos da solução correta, mas havia algo que não batia certo. Tentámos várias variações, sem sucesso. Foi frustrante. Depois de alguns meses, ficámos bloqueados. Não conseguíamos vislumbrar uma solução para o problema, e por isso

deixámo-lo de lado durante algum tempo para trabalhar noutras questões.

No final de fevereiro de 2016, estava no meu escritório à espera da minha esposa, Janet, para almoçarmos. Tinha uma chávena da Numenta na mão e reparei nos meus dedos a tocar-lhe. Fiz-me uma pergunta simples: O que é que o meu cérebro precisa de saber para prever o que os meus dedos sentirão ao moverem-se? Se um dos meus dedos está de lado da chávena e se move na direção do topo, o meu cérebro prevê que sentirei a curva arredondada do rebordo. O meu cérebro faz esta previsão antes do dedo tocar no rebordo. O que é que o cérebro precisa de saber para fazer essa previsão? A resposta é fácil de enunciar: o cérebro precisa de saber duas coisas — qual é o objeto que está a ser tocado (neste caso, a chávena) e onde o meu dedo estará na chávena depois do movimento.

Repare que o cérebro precisa de saber onde está o meu dedo em relação à chávena. Não importa onde o meu dedo está em relação ao meu corpo, nem importa a posição ou orientação da chávena. A chávena pode estar inclinada para a esquerda ou para a direita. Pode estar à minha frente ou de lado. O que importa é a localização do meu dedo em relação à chávena.

Esta observação significa que devem existir neurónios no neocórtex que representam a localização do meu dedo num referencial fixo ao objeto. O sinal relacionado com o movimento que procurávamos — o sinal de que precisávamos para prever o próximo estímulo — era: *"localização no objeto"*.

Provavelmente aprendeu sobre referenciais espaciais no liceu. Os eixos x, y e z, que definem a localização de algo no espaço, são um exemplo de referencial. Outro exemplo familiar são a latitude e a longitude, que definem posições na superfície da Terra. No início, foi difícil para nós imaginar como é que neurónios poderiam representar algo como coordenadas x, y e z. Mas ainda mais intrigante era que os neurónios pudessem atribuir um referencial a um objeto, como uma chávena de café. O referencial da chávena é relativo à própria chávena; por isso, o referencial tem de se mover com ela.

Imagine uma cadeira de escritório. O meu cérebro prevê o que irei sentir ao tocar na cadeira, tal como prevê o que sentirei ao tocar na chávena. Logo, devem existir neurónios no meu neocórtex que conhecem a localização do meu dedo em relação à cadeira, o que significa que o meu neocórtex tem de estabelecer um referencial fixo à cadeira. Se eu rodar a cadeira em círculo, o referencial roda com ela. Se eu virar a cadeira ao contrário, o referencial inverte-se. Pode pensar no referencial como uma grelha invisível tridimensional que envolve e está ligada ao objeto (neste caso, à cadeira). Os neurónios são estruturas simples. Custava-nos imaginar que fossem capazes de criar e atribuir referenciais a objetos, mesmo quando esses objetos se movem e rodam no mundo exterior. Mas a surpresa ainda foi maior.

Diferentes partes do meu corpo (pontas dos dedos, palma da mão, lábios) podem tocar na chávena ao mesmo tempo. Cada parte do corpo que toca na chávena faz uma previsão separada sobre o que sentirá, com base na sua localização específica na chávena. Ou seja, o cérebro não faz apenas uma previsão; faz dezenas ou até centenas de previsões ao mesmo tempo. O neocórtex tem de saber a localização, relativamente à chávena, de cada parte do corpo que a está a tocar.

Dei-me conta de que a visão faz a mesma coisa que o tato. Pedaços da retina são análogos a pedaços de pele. Cada área da sua retina vê apenas uma pequena parte de um objeto, do mesmo modo que cada área da pele toca apenas uma pequena parte do mesmo. O cérebro não processa uma imagem como um todo; começa com a imagem na parte de trás do olho, mas decompõe essa imagem em centenas de fragmentos. Depois, atribui a cada fragmento uma localização relativa ao objeto que está a ser observado.

Criar referenciais e acompanhar localizações não é uma tarefa trivial. Eu sabia que isso exigiria vários tipos diferentes de neurónios e múltiplas camadas de células para realizar tais cálculos. Uma vez que os circuitos complexos em cada coluna cortical são semelhantes entre si, as localizações e os referenciais têm de ser propriedades universais do neocórtex. Cada coluna do neocórtex — quer represente entrada visual, tátil, auditiva, linguagem ou pensamento

de ordem superior — tem de conter neurónios que representem referenciais e localizações.

Até então, a maioria dos neurocientistas, eu incluído, pensava que o neocórtex servia sobretudo para processar informação sensorial. O que percebi naquele dia foi que devemos pensar no neocórtex como um sistema que processa principalmente referenciais espaciais. A maior parte da sua circuitaria existe para criar referenciais e seguir localizações. A entrada sensorial, naturalmente, é essencial. Como explicarei nos capítulos seguintes, o cérebro constrói modelos do mundo associando a informação sensorial a localizações em referenciais.

Por que são os referenciais tão importantes? O que ganha o cérebro por tê-los? Em primeiro lugar, um referencial permite ao cérebro aprender a estrutura de algo. Uma chávena de café é um objeto porque é composta por um conjunto de características e superfícies dispostas em relação umas às outras no espaço. Do mesmo modo, uma cara é constituída por um nariz, olhos e boca organizados em posições relativas. É necessário um referencial para especificar essas posições relativas e a estrutura dos objetos.

Em segundo lugar, ao definir um objeto através de um referencial, o cérebro pode manipular o objeto inteiro de uma só vez. Por exemplo, um carro tem muitas características organizadas relativamente entre si. Uma vez aprendido o conceito de "carro", conseguimos imaginar como seria visto de diferentes ângulos ou se fosse esticado numa determinada dimensão. Para realizar tais feitos, o cérebro só precisa de rodar ou esticar o referencial, e todas as características do carro rodam ou esticam com ele.

Em terceiro lugar, um referencial é necessário para planear e realizar movimentos. Suponhamos que o meu dedo está a tocar na parte frontal do meu telemóvel e quero carregar no botão de ligar/desligar no topo. Se o meu cérebro sabe a localização atual do dedo e a localização do botão, então consegue calcular o movimento necessário para levar o dedo do ponto atual até ao novo ponto desejado. É necessário um referencial relativo ao telemóvel para efetuar esse cálculo.

Referenciais são utilizados em muitos campos. Os roboticistas dependem deles para planear os movimentos do braço ou do corpo

de um robô. Também são usados em filmes de animação, para renderizar personagens enquanto se movimentam. Algumas pessoas já tinham sugerido que os referenciais poderiam ser necessários para certas aplicações de IA. Mas, tanto quanto sei, nunca houve uma discussão significativa sobre a hipótese de o neocórtex trabalhar com referenciais, nem de que a função da maioria dos neurónios em cada coluna cortical seria criar referenciais e seguir localizações. Agora, isso parece-me óbvio.

Vernon Mountcastle argumentou que existia um algoritmo universal presente em cada coluna cortical, mas não sabia qual era esse algoritmo. Francis Crick escreveu que necessitávamos de uma nova estrutura conceptual para compreender o cérebro, e também ele não sabia qual deveria ser essa estrutura. Naquele dia de 2016, com a chávena na mão, percebi que o algoritmo de Mountcastle e a estrutura de Crick estavam ambos baseados em referenciais. Ainda não compreendia como é que os neurónios conseguiam fazer isso, mas sabia que tinha de ser verdade. Os referenciais eram o ingrediente em falta, a chave para desvendar o mistério do neocórtex e para compreender a inteligência.

Todas estas ideias sobre localizações e referenciais surgiram-me como que num segundo. Fiquei tão entusiasmado que saltei da cadeira e corri para contar ao meu colega Subutai Ahmad. Na corrida dos poucos metros até à sua secretária, cruzei-me com a Janet e quase a derrubei. Estava ansioso por falar com o Subutai, mas enquanto ajudava a Janet a recuperar o equilíbrio e lhe pedia desculpa, percebi que seria mais sensato falar com ele mais tarde. A Janet e eu conversámos sobre referenciais e localizações enquanto partilhávamos um iogurte gelado.

> **Este é um bom momento para abordar uma pergunta que me fazem frequentemente: como posso falar com confiança sobre uma teoria que ainda não foi testada experimentalmente? Acabei de descrever uma dessas situações. Tive a intuição de que o neocórtex está imbuído de referenciais, e comecei de imediato a falar sobre isso com convicção. No momento em que escrevo este livro, há cada vez mais evidências que apoiam esta nova ideia, mas ela ainda não foi testada de forma exaustiva. E, no entanto, não hesito em descrever esta ideia como um facto. Eis porquê.**

À medida que trabalhamos sobre um problema, vamos descobrindo aquilo a que chamo restrições. Restrições são elementos que a solução do problema tem de respeitar. Dei alguns exemplos de restrições ao descrever a memória de sequências, como por exemplo o requisito do Nomeie essa música. A anatomia e a fisiologia do cérebro são também restrições. A teoria do cérebro tem de explicar, em última análise, todos os detalhes do cérebro, e uma teoria correta não pode violar nenhum desses detalhes.

Quanto mais tempo se trabalha num problema, mais restrições se descobrem, e mais difícil se torna imaginar uma solução. Os momentos "eureka" que descrevi neste capítulo referem-se a problemas em que trabalhámos durante anos. Por isso, compreendíamos profundamente esses problemas e a nossa lista de restrições era longa. A probabilidade de uma solução estar correta aumenta exponencialmente com o número de restrições que satisfaz. É como resolver um puzzle de palavras cruzadas: há muitas palavras que podem corresponder a uma pista isolada. Se escolher uma delas, pode estar errada. Mas se encontrar duas palavras que se cruzam e que funcionam, então é muito mais provável que ambas estejam corretas. Se encontrar dez palavras que se cruzam corretamente, a hipótese de estarem todas erradas é mínima. Pode-se escrever a resposta a tinta, sem receios.

Os momentos "eureka" surgem quando uma nova ideia satisfaz múltiplas restrições. Quanto mais tempo se trabalhou sobre um problema — e, por conseguinte, quanto mais restrições a solução resolve — maior é a sensação de clareza e mais confiança se tem na resposta encontrada. A ideia de que o neocórtex está imbuído de referenciais resolveu tantas restrições que soube de imediato que era correta.

Levou-nos mais de três anos a desenvolver as implicações desta descoberta e, no momento em que escrevo, ainda não terminámos. Já publicámos vários artigos sobre o tema. O primeiro intitula-se "*Uma Teoria de Como as Colunas no Neocórtex Permitem Aprender a Estrutura do Mundo*". Este artigo parte do mesmo circuito que descrevemos no artigo de 2016 sobre neurónios e

memória de sequências. Depois, acrescentámos uma camada de neurónios que representa a localização e uma segunda camada que representa o objeto que está a ser percecionado. Com estas adições, demonstrámos que uma única coluna cortical pode aprender a forma tridimensional dos objetos, através do processo de sentir e mover, sentir e mover.

Por exemplo, imagine que mete a mão numa caixa preta e toca num objeto novo com um dedo. Pode aprender a forma do objeto inteiro movendo o dedo pelas suas extremidades. O nosso artigo explicava como uma única coluna cortical pode fazer isto. Mostrámos também como uma coluna pode reconhecer um objeto anteriormente aprendido do mesmo modo, por exemplo, ao mover um dedo. Em seguida, mostrámos como múltiplas colunas no neocórtex colaboram para reconhecer objetos de forma mais rápida. Por exemplo, se meter a mão na caixa preta e agarrar um objeto desconhecido com a mão inteira, poderá reconhecê-lo com menos movimentos e, em certos casos, com um único gesto.

Estávamos nervosos quanto à submissão deste artigo e debatemos se deveríamos esperar. Estávamos a propor que todo o neocórtex funciona criando referenciais, com muitos milhares deles ativos em simultâneo. Era uma ideia radical. E, no entanto, não tínhamos ainda uma proposta de como os neurónios criam efetivamente os referenciais. O nosso argumento era algo como: "*Deduzirmos que as localizações e os referenciais têm de existir e, assumindo que existem, eis como uma coluna cortical pode funcionar. E, ah, já agora, não sabemos como é que os neurónios conseguem realmente criar referenciais.*" Decidimos submeter o artigo na mesma. Perguntei a mim próprio: Gostaria eu de ler este artigo, mesmo estando incompleto? A minha resposta foi sim. A ideia de que o neocórtex representa localizações e referenciais em cada coluna era demasiado entusiasmante para ser adiada só porque não sabíamos ainda como os neurónios o faziam. Estava confiante de que a ideia base era correta.

Leva-se muito tempo a compor um artigo científico. Só o texto pode demorar meses a escrever, e normalmente há simulações para realizar, o que pode acrescentar ainda mais meses ao processo. Perto do fim deste trabalho, tive uma ideia que

acrescentámos ao artigo mesmo antes da submissão. Sugeri que talvez encontrássemos a resposta para como os neurónios no neocórtex criam referenciais olhando para uma parte mais antiga do cérebro chamada córtex entorrinal. Quando o artigo foi aceite, alguns meses depois, já sabíamos que esta conjetura estava correta, como irei explicar no próximo capítulo.

Acabámos de percorrer bastante terreno, por isso vamos fazer uma breve recapitulação. O objetivo deste capítulo foi apresentar-lhe a ideia de que cada coluna cortical no neocórtex cria referenciais. Levei-lhe através dos passos que demos para chegar a esta conclusão. Começámos com a ideia de que o neocórtex aprende um modelo rico e detalhado do mundo, que utiliza para prever constantemente quais serão os seus próximos estímulos sensoriais. Depois perguntámos como é que os neurónios conseguem fazer essas previsões. Isso levou-nos a uma nova teoria segundo a qual a maioria das previsões é representada por picos dendríticos que alteram temporariamente a voltagem no interior de um neurónio e fazem com que este dispare ligeiramente mais cedo do que dispararia de outro modo. As previsões não são transmitidas através do axónio da célula para outros neurónios, o que explica por que é que não temos consciência da maior parte delas. Em seguida, mostrámos como os circuitos no neocórtex que utilizam este novo modelo de neurónio conseguem aprender e prever sequências. Aplicámos esta ideia à questão de como é que um circuito deste tipo poderia prever o próximo estímulo sensorial quando os estímulos estão a mudar devido aos nossos próprios movimentos. Para fazer estas previsões sensório-motoras, deduzimos que cada coluna cortical tem de saber a localização do seu estímulo em relação ao objeto que está a ser percecionado. Para isso, uma coluna cortical necessita de um referencial fixo ao objeto.

CAPÍTULO 5

Mapas no Cérebro

Demorámos anos a deduzir que os referenciais espaciais existem por todo o neocórtex, mas, em retrospetiva, poderíamos tê-lo compreendido há muito tempo através de uma simples observação. Neste momento, estou sentado numa pequena zona de estar do escritório da Numenta. Perto de mim há três cadeiras confortáveis, semelhantes àquela onde me encontro. Para além das cadeiras, há várias secretárias independentes. Para além das secretárias, vejo o antigo tribunal do condado do outro lado da rua. A luz proveniente destes objetos entra nos meus olhos e é projetada na retina. As células da retina convertem a luz em impulsos elétricos. É aqui que a visão começa, na parte de trás do olho. Por que é, então, que não percebemos os objetos como estando dentro do olho? Se as cadeiras, as secretárias e o tribunal estão representados lado a lado na minha retina, como é que os perceciono como estando a diferentes distâncias e em diferentes localizações? Do mesmo modo, se ouço um carro a aproximar-se, por que o perceciono como estando a trinta metros à minha direita, e não dentro do ouvido, onde o som efetivamente é captado?

Esta simples observação — de que percebemos os objetos como estando em algum lugar, e não nos olhos ou ouvidos, mas sim numa determinada localização no mundo — indica-nos que o cérebro tem de possuir neurónios cuja atividade representa a localização de cada objeto que percecionamos.

No final do capítulo anterior, referi que estávamos receosos de submeter o nosso primeiro artigo sobre referenciais espaciais, porque, nessa altura, não sabíamos como é que os neurónios do neocórtex conseguiam fazer isso. Estávamos a propor uma nova teoria importante sobre o funcionamento do neocórtex, mas a teoria baseava-se essencialmente em deduções lógicas. O artigo seria mais sólido se conseguíssemos mostrar como é que os neurónios

realizam essa função. No dia anterior à submissão, adicionei algumas linhas de texto sugerindo que a resposta poderia ser encontrada numa parte mais antiga do cérebro chamada córtex entorrinal. Vou agora explicar por que fizemos essa sugestão com uma história sobre evolução.

Um Conto Evolutivo

Quando os animais começaram a deslocar-se pelo mundo, necessitavam de um mecanismo que lhes permitisse decidir para onde se mover. Animais simples possuem mecanismos simples. Por exemplo, algumas bactérias seguem gradientes. Se a quantidade de um recurso necessário, como alimento, estiver a aumentar, então é mais provável que continuem a mover-se na mesma direção. Se a quantidade estiver a diminuir, então é mais provável que mudem de direção e tentem outro caminho. Uma bactéria não sabe onde está; não possui qualquer meio de representar a sua localização no mundo. Apenas avança e usa uma regra simples para decidir quando deve virar. Um animal ligeiramente mais sofisticado, como uma minhoca, pode mover-se para permanecer dentro de intervalos desejáveis de calor, alimento e água, mas também não sabe onde está no jardim. Não sabe quão longe está do caminho de tijolos, nem a direção ou a distância até ao poste de vedação mais próximo.

Agora considere as vantagens oferecidas a um animal que sabe onde está — um animal que sabe sempre a sua localização em relação ao ambiente. Esse animal pode lembrar-se de onde encontrou alimento no passado e dos locais que utilizou como abrigo. Pode então calcular como chegar do local onde se encontra até esses, e outros, locais visitados anteriormente. Pode recordar o caminho que percorreu até ao bebedouro e o que aconteceu em diferentes pontos ao longo do trajeto. Saber onde se está, e onde estão outras coisas no mundo, traz muitas vantagens — mas requer um referencial.

Recorde que um referencial é como a grelha de um mapa. Por exemplo, num mapa de papel, poderá localizar algo usando linhas e colunas etiquetadas, como a linha D e a coluna 7. As linhas e colunas de um mapa constituem o referencial para a área

representada. Se um animal tiver um referencial para o seu mundo, então, à medida que o explora, pode registar o que encontrou em cada localização. Quando quiser chegar a um local específico — como um abrigo — pode usar esse referencial para calcular como chegar lá a partir do ponto onde se encontra. Ter um referencial para o mundo é útil para a sobrevivência.

Ser capaz de se orientar no mundo é tão valioso que a evolução descobriu múltiplos métodos para o fazer. Por exemplo, algumas abelhas conseguem comunicar distâncias e direções através de uma forma de dança. Mamíferos, como nós, possuem um poderoso sistema de navegação interno. Existem neurónios numa parte antiga do nosso cérebro que são conhecidos por construir mapas dos lugares que visitámos, e estes neurónios estiveram sob pressão evolutiva durante tanto tempo que estão afinados para cumprir essa função com precisão. Nos mamíferos, as partes antigas do cérebro onde estes neurónios criadores de mapas se encontram chamam-se hipocampo e córtex entorrinal. No ser humano, estes órgãos têm aproximadamente o tamanho de um dedo. Existe um conjunto de cada lado do cérebro, perto do centro.

Mapas no Cérebro Antigo

Em 1971, o cientista John O'Keefe e o seu aluno Jonathan Dostrovsky colocaram um fio no cérebro de um rato. Esse fio registava a atividade elétrica (picos de atividade) de um único neurónio no hipocampo. O fio era dirigido para cima, em direção ao teto, de forma a que pudessem registar a atividade da célula enquanto o rato se movia e explorava o seu ambiente — normalmente uma grande caixa sobre uma mesa. Descobriram aquilo a que hoje chamamos *células de lugar*: neurónios que disparam sempre que o rato se encontra numa determinada localização de um ambiente específico. Uma célula de lugar funciona como um marcador de "você está aqui" num mapa. À medida que o rato se desloca, diferentes células de lugar ativam-se em cada nova localização. Se o rato regressa a um local onde já esteve, a mesma célula de lugar volta a ativar-se.

Em 2005, cientistas no laboratório de May-Britt Moser e Edvard Moser utilizaram uma configuração experimental semelhante,

novamente com ratos. Nas suas experiências, registaram sinais de neurónios no córtex entorrinal, adjacente ao hipocampo. Descobriram aquilo a que hoje chamamos *células de grelha*, que disparam em múltiplas localizações dentro de um ambiente. As localizações onde uma célula de grelha se ativa formam um padrão em grelha. Se o rato se move em linha reta, a mesma célula de grelha ativa-se repetidamente, a intervalos igualmente espaçados.

Os detalhes sobre como funcionam as *células de lugar* e as *células de grelha* são complexos e ainda não totalmente compreendidos, mas podemos considerá-las como formando um mapa do ambiente ocupado pelo rato. As células de grelha são como as linhas e colunas de um mapa de papel, sobrepostas ao ambiente do animal. Permitem ao animal saber onde está, prever onde estará ao mover-se e planear movimentos. Por exemplo, se estou na localização B4 de um mapa e quero chegar à localização D6, posso usar a grelha do mapa para saber que tenho de me mover dois quadrados para a direita e dois para baixo.

Mas as células de grelha, por si só, não dizem o que existe numa determinada localização. Por exemplo, se eu lhe disser que está na localização A6 de um mapa, essa informação não lhe diz o que encontrará nesse lugar. Para saber o que há em A6, tem de olhar para o mapa e ver o que está impresso nesse quadrado. As células de lugar são como os detalhes impressos nesse quadrado. As células de lugar que se ativam dependem daquilo que o rato sente numa localização específica. As células de lugar dizem ao rato onde está, com base nos estímulos sensoriais, mas não são úteis para planear movimentos — isso requer as células de grelha. Os dois tipos de células trabalham em conjunto para criar um modelo completo do ambiente do rato.

Sempre que um rato entra num ambiente, as células de grelha estabelecem uma moldura de referência. Se o ambiente for novo, as células de grelha criam uma nova moldura de referência. Se o rato reconhecer o ambiente, as células de grelha restabelecem a moldura de referência previamente utilizada. Este processo é análogo ao de entrar numa cidade. Se olhar em redor e perceber que já esteve ali antes, pega no mapa correto dessa cidade. Se a cidade lhe parecer desconhecida, então pega numa folha em branco

e começa a criar um novo mapa. À medida que passeia pela cidade, anota no seu mapa o que vê em cada localização. É isso que fazem as células de grelha e as células de lugar. Criam mapas únicos para cada ambiente. À medida que o rato se move, as células de grelha e as células de lugar ativas mudam para refletir a nova localização.

Os humanos também possuem células de grelha e células de lugar. A menos que esteja completamente desorientado, tem sempre uma noção de onde está. Neste momento, estou de pé no meu escritório. Mesmo que feche os olhos, a minha perceção do local onde estou mantém-se, e continuo a saber onde me encontro. Com os olhos fechados, dou dois passos para a direita e a minha sensação de localização na sala altera-se. As células de grelha e as células de lugar no meu cérebro criaram um mapa do meu escritório, e acompanham a minha posição nele, mesmo quando os olhos estão fechados. À medida que caminho, as células ativas vão mudando para refletir a minha nova localização. Humanos, ratos, na verdade todos os mamíferos, utilizam o mesmo mecanismo para saber onde estão. Todos temos células de grelha e células de lugar que constroem modelos dos locais por onde passámos.

Mapas no Novo Cérebro

Quando estávamos a escrever o nosso artigo de 2017 sobre localizações e molduras de referência no neocórtex, eu já tinha algum conhecimento sobre células de lugar e células de grelha. Ocorreu-me que saber a localização do meu dedo em relação a uma chávena de café é semelhante a saber a localização do meu corpo em relação a uma sala. O meu dedo move-se ao redor da chávena da mesma forma que o meu corpo se move numa divisão. Percebi então que o neocórtex poderia conter neurónios equivalentes aos que existem no hipocampo e no córtex entorrinal. Estas células de lugar corticais e células de grelha corticais aprenderiam modelos de objetos de maneira semelhante à forma como as células de lugar e as células de grelha do cérebro antigo aprendem modelos de ambientes.

Dado o seu papel na navegação básica, é quase certo que as células de lugar e as células de grelha são evolutivamente mais antigas do que o neocórtex. Por isso, presumi que seria mais

provável que o neocórtex criasse molduras de referência utilizando um derivado das células de grelha, do que ter evoluído um novo mecanismo desde o início. Contudo, em 2017, não estávamos cientes de qualquer evidência de que o neocórtex possuísse algo semelhante a células de grelha ou células de lugar — era uma especulação fundamentada.

Pouco depois de o nosso artigo de 2017 ter sido aceite, tomámos conhecimento de experiências recentes que sugeriam que poderiam existir células de grelha em partes do neocórtex. (Falarei dessas experiências no Capítulo 7.) Isso foi encorajador. Quanto mais estudávamos a literatura relacionada com células de grelha e células de lugar, mais confiantes ficávamos de que células com funções semelhantes existiam em cada coluna cortical. Apresentámos esse argumento pela primeira vez num artigo de 2019, intitulado "*Uma Estrutura para a Inteligência e a Função Cortical com Base em Células de Grelha no Neocórtex*".

Mais uma vez, para aprender um modelo completo de algo, são necessárias tanto células de grelha como células de lugar. As células de grelha criam uma moldura de referência para especificar localizações e planear movimentos. Mas também é necessário o conteúdo sensorial, representado pelas células de lugar, para associar os dados sensoriais às localizações na moldura de referência.

Os mecanismos de mapeamento no neocórtex não são uma cópia exata dos existentes no cérebro antigo. A evidência sugere que o neocórtex utiliza os mesmos mecanismos neuronais básicos, mas de formas diferentes. É como se a natureza tivesse reduzido o hipocampo e o córtex entorrinal à sua forma mínima, feito dezenas de milhares de cópias e disposto essas cópias lado a lado em colunas corticais. Foi assim que surgiu o neocórtex.

As células de grelha e de lugar no cérebro antigo acompanham sobretudo a localização de uma coisa: o corpo. Elas sabem onde o corpo está no ambiente atual. O neocórtex, por sua vez, possui cerca de 150.000 cópias deste circuito, uma por cada coluna cortical. Assim, o neocórtex acompanha milhares de localizações em simultâneo. Por exemplo, cada pequena porção da sua pele e cada pequena porção da sua retina tem a sua própria moldura de

referência no neocórtex. As suas cinco pontas dos dedos a tocar numa chávena são como cinco ratos a explorar uma caixa.

Mapas Enormes em Espaços Minúsculos

Então, como é que se apresenta um modelo no cérebro? Como é que o neocórtex consegue armazenar centenas de modelos em cada milímetro quadrado? Para compreender como isto funciona, vamos voltar à analogia do mapa em papel. Imaginemos que tenho um mapa de uma cidade. Estendo-o sobre uma mesa e vejo que está marcado com linhas e colunas que o dividem em cem quadrados. A1 é o canto superior esquerdo e J10 é o canto inferior direito. Impresso em cada quadrado estão elementos que posso encontrar nessa parte da cidade. Pego numa tesoura e recorto cada quadrado, assinalando-o com as suas coordenadas de grelha: B6, G1, etc. Assinalo também cada quadrado com o nome Cidade 1. Faço o mesmo com mais nove mapas, cada um representando uma cidade diferente. Fico agora com mil quadrados: cem quadrados de mapa para cada uma das dez cidades. Baralho os quadrados e empilho-os. Apesar de a minha pilha conter dez mapas completos, apenas uma localização pode ser vista de cada vez. Agora, alguém coloca-me uma venda nos olhos e larga-me num local aleatório de uma das dez cidades.

Tiro a venda e olho em volta. Ao início, não sei onde estou. Depois vejo que estou em frente a uma fonte com uma escultura de uma mulher a ler um livro. Folheio os meus quadrados de mapa, um de cada vez, até encontrar um que mostre essa fonte. Esse quadrado está rotulado como Cidade 3, localização D2. Agora sei em que cidade estou e onde estou nessa cidade.

Posso fazer várias coisas a seguir. Por exemplo, posso prever o que verei se começar a andar. A minha localização atual é D2. Se caminhar para leste, estarei em D3. Procuro na minha pilha o quadrado rotulado como Cidade 3, D3. Mostra um parque infantil. Desta forma, posso prever o que irei encontrar se me mover numa determinada direção.

Talvez queira ir até à biblioteca da cidade. Posso procurar na minha pilha até encontrar um quadrado que mostre uma biblioteca

na Cidade 3. Esse quadrado está rotulado como G7. Sabendo que estou em D2, posso calcular que tenho de caminhar três quadrados para leste e cinco quadrados para sul para chegar à biblioteca. Posso escolher diferentes percursos para lá chegar. Usando os meus quadrados de mapa, um de cada vez, posso visualizar o que encontrarei ao longo de cada trajeto possível. Escolho um que me leve a passar por uma geladaria.

Agora considere-se um cenário diferente. Depois de ser largado num local desconhecido e de tirar a venda, vejo uma cafetaria. Mas, ao olhar para a minha pilha de quadrados, encontro cinco que mostram uma cafetaria semelhante. Duas estão na mesma cidade, e as outras três em cidades diferentes. Posso estar em qualquer um desses cinco locais. O que devo fazer? Posso eliminar a ambiguidade movendo-me. Olho para os cinco quadrados onde posso estar e depois consulto o que veria se andasse para sul a partir de cada um deles. A resposta é diferente para cada um dos cinco. Para descobrir onde estou, então, caminho fisicamente para sul. O que encontro elimina a minha incerteza. Agora sei onde estou.

Esta forma de usar mapas é diferente da forma como os usamos habitualmente. Primeiro, a nossa pilha de quadrados de mapa contém todos os nossos mapas. Desta forma, usamos a pilha para descobrir tanto em que cidade estamos como onde estamos nessa cidade.

Em segundo lugar, se estivermos incertos quanto ao local onde nos encontramos, então podemos determinar a nossa cidade e localização através do movimento. É isso que acontece quando metemos a mão numa caixa escura e tocamos num objeto desconhecido com um dedo. Com um único toque, provavelmente não conseguimos determinar que objeto estamos a sentir. Pode ser necessário mover o dedo uma ou mais vezes para fazer essa determinação. Ao movermo-nos, descobrimos duas coisas em simultâneo: no momento em que reconhecemos que objeto estamos a tocar, também sabemos onde está o nosso dedo nesse objeto.

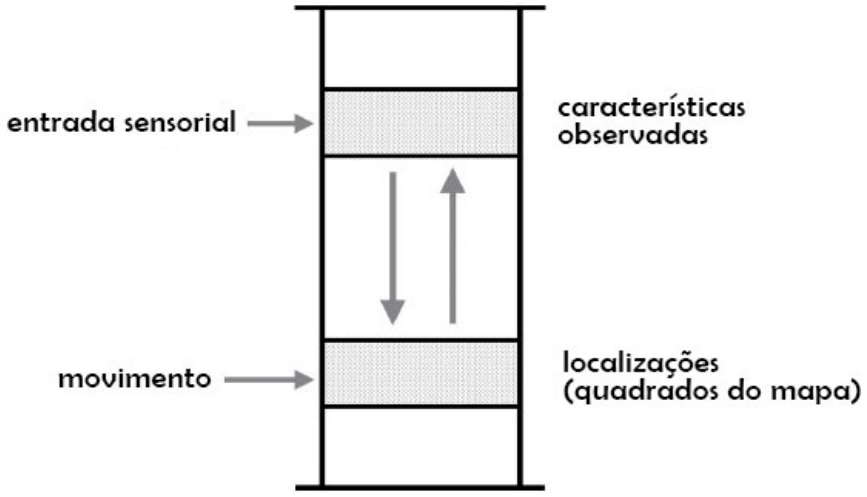
Por fim, este sistema pode ser escalado para lidar com um grande número de mapas e fazê-lo rapidamente. Na analogia do

mapa em papel, descrevi a consulta dos quadrados de mapa um a um. Isso poderia demorar muito tempo se tivéssemos muitos mapas. No entanto, os neurónios utilizam o que se chama memória associativa. Os detalhes não são importantes aqui, mas essa memória permite que os neurónios pesquisem todos os quadrados de mapa de uma só vez. Os neurónios demoram o mesmo tempo a pesquisar mil mapas que a pesquisar apenas um.

Mapas Numa Coluna Cortical

Vamos agora considerar como é que modelos em forma de mapa são implementados por neurónios no neocórtex. A nossa teoria afirma que cada coluna cortical pode aprender modelos de objetos completos. Portanto, cada coluna — cada milímetro quadrado do neocórtex — possui o seu próprio conjunto de quadrados de mapa. A forma como uma coluna cortical faz isto é complexa, e ainda não a compreendemos totalmente, mas conhecemos os princípios básicos.

Recordemos que uma coluna cortical possui múltiplas camadas de neurónios. Várias dessas camadas são necessárias para criar os quadrados de mapa. Eis um diagrama simplificado para dar uma ideia do que pensamos estar a acontecer numa coluna cortical.



Um modelo de uma coluna cortical

Esta figura representa duas camadas de neurónios (os blocos sombreados) numa coluna cortical. Embora uma coluna seja

minúscula, com cerca de um milímetro de largura, cada uma destas camadas pode ter dez mil neurónios.

A camada superior recebe o input sensorial da coluna. Quando um estímulo chega, leva à ativação de várias centenas de neurónios. Na analogia do mapa em papel, a camada superior representa aquilo que se observa num determinado local, como por exemplo a fonte.

A camada inferior representa a localização atual num referencial. Na analogia, a camada inferior representa uma localização — por exemplo, Cidade 3, D2 — mas não representa o que é observado nesse local. É como um quadrado em branco, rotulado apenas com "Cidade 3, localização D2".

As duas setas verticais representam as ligações entre os quadrados de mapa em branco (a camada inferior) e o que é observado nesse local (a camada superior). A seta descendente representa a forma como uma característica observada, como a fonte, é associada a uma determinada localização numa determinada cidade. A seta ascendente associa uma localização específica — Cidade 3, D2 — a uma característica observada.

A camada superior é aproximadamente equivalente às células de lugar, e a camada inferior é aproximadamente equivalente às células de grelha.

Aprender um novo objeto, como uma chávena de café, é em grande parte conseguido através da aprendizagem das ligações entre as duas camadas, ou seja, das setas verticais. Por outras palavras, um objeto como uma chávena de café é definido por um conjunto de características observadas (camada superior) associadas a um conjunto de localizações na chávena (camada inferior). Se se conhece a característica, então pode-se determinar a localização. Se se conhece a localização, pode-se prever a característica.

O fluxo básico da informação ocorre da seguinte forma: Um input sensorial chega e é representado pelos neurónios na camada superior. Isto invoca a localização na camada inferior que está associada ao estímulo. Quando ocorre movimento, como mover um dedo, a camada inferior muda para a nova localização esperada, o que gera uma previsão do próximo input na camada superior.

Se o input original for ambíguo, como no caso do café, a rede ativa múltiplas localizações na camada inferior — por exemplo, todas as localizações onde existe um café. Isto é o que acontece quando se toca na borda de uma chávena de café com um dedo. Muitos objetos têm uma borda, por isso inicialmente não se consegue ter a certeza de que objeto se está a tocar. Quando se movimenta, a camada inferior atualiza todas as localizações possíveis, o que gera múltiplas previsões na camada superior. O input seguinte eliminará quaisquer localizações que não correspondam.

Simulámos este circuito de duas camadas em software, utilizando pressupostos realistas quanto ao número de neurónios em cada camada. As nossas simulações mostraram que não só colunas corticais individuais conseguem aprender modelos de objetos, como cada coluna pode aprender centenas deles. O mecanismo neuronal e as simulações estão descritos no nosso artigo de 2019, "Localizações no Neocórtex: Uma Teoria de Reconhecimento de Objetos Sensório-Motores com recurso a Células da Grelha Cortical."

Orientação

Há outras funções que uma coluna cortical tem de desempenhar para aprender modelos de objetos. Por exemplo, é necessária uma representação da orientação. Suponha que sabe em que cidade está e qual é a sua localização nessa cidade. Agora pergunto-lhe: «O que verá se andar um quarteirão para a frente?» Você responderia: «Para que lado estou a caminhar?» Saber a sua localização não é suficiente para prever o que verá ao mover-se; é também necessário saber para que lado está voltado, ou seja, a sua orientação. A orientação também é essencial para prever o que se verá a partir de uma determinada localização. Por exemplo, estando numa esquina, poderá ver uma biblioteca se estiver virado para norte e um parque infantil se estiver virado para sul.

Existem neurónios no cérebro antigo chamados *células de direção da cabeça*. Como o nome indica, estas células representam a direção para a qual a cabeça do animal está orientada. As células de direção da cabeça funcionam como uma bússola, mas não estão

alinhadas com o norte magnético; estão alinhadas com uma sala ou ambiente específico. Se estiver numa sala familiar e fechar os olhos, mantém a noção de para que lado está voltado. Se rodar o corpo com os olhos fechados, a sua perceção da direção muda. Essa sensação é criada pelas suas células de direção da cabeça. Quando roda o corpo, essas células mudam para refletir a sua nova orientação na sala.

As colunas corticais têm de possuir células que desempenhem uma função equivalente à das células de direção da cabeça. Chamamos-lhes, de forma mais genérica, *células de orientação*. Imagine que está a tocar na borda de uma chávena de café com o seu dedo indicador. A impressão que o dedo recebe depende da sua orientação. Pode, por exemplo, manter o dedo na mesma localização, mas rodá-lo em torno do ponto de contato. À medida que o faz, a sensação no dedo muda. Portanto, para poder prever o seu input, uma coluna cortical tem de possuir uma representação da orientação. Para simplificar, não mostrei as células de orientação nem outros pormenores no diagrama anterior de uma coluna cortical.

Para resumir, propusemos que cada coluna cortical aprende modelos de objetos. As colunas fazem-no usando o mesmo método básico que o cérebro antigo utiliza para aprender modelos de ambientes. Propusemos, por conseguinte, que cada coluna cortical possui um conjunto de células equivalentes às células de grelha, outro conjunto equivalente às células de lugar, e ainda outro conjunto equivalente às células de direção da cabeça — todas elas inicialmente descobertas em partes do cérebro antigo. Chegámos a esta hipótese por dedução lógica. No Capítulo 7, apresentarei as evidências experimentais crescentes que sustentam a nossa proposta.

Mas, antes disso, vamos voltar a nossa atenção para o neocórtex como um todo. Recorde que cada coluna cortical é pequena, com a largura de um fio de esparguete fino, e que o neocórtex é grande, do tamanho de um guardanapo de jantar. Existem, portanto, cerca de 150.000 colunas no neocórtex humano. Nem todas essas colunas estão a modelar objetos. O que fazem as restantes colunas é o tema do próximo capítulo.

CAPÍTULO 6

Conceitos, Linguagem e Pensamento de Alto Nível

As nossas funções cognitivas superiores são aquilo que mais nos distingue dos nossos primos primatas. A nossa capacidade de ver e ouvir é semelhante à de um macaco, mas só os humanos usam linguagem complexa, constroem ferramentas sofisticadas como computadores, e são capazes de raciocinar sobre conceitos como evolução, genética e democracia.

Vernon Mountcastle propôs que cada coluna do neocórtex desempenha a mesma função básica. Para que isto seja verdade, então a linguagem e outras capacidades cognitivas de alto nível são, num nível fundamental, iguais a ver, tocar e ouvir. Isto não é evidente. Ler Shakespeare não parece semelhante a pegar numa chávena de café, mas essa é a implicação da proposta de Mountcastle.

Mountcastle sabia que as colunas corticais não são completamente idênticas. Existem diferenças físicas, por exemplo, entre colunas que recebem input dos seus dedos e colunas que compreendem linguagem, mas há mais semelhanças do que diferenças. Por conseguinte, Mountcastle deduziu que deve existir alguma função básica que está na base de tudo o que o neocórtex faz — não apenas a perceção, mas todas as coisas que associamos à inteligência.

A ideia de que capacidades tão diversas como visão, tato, linguagem e filosofia são, fundamentalmente, a mesma coisa, é difícil de aceitar para muitas pessoas. Mountcastle não propôs qual seria essa função comum, e é difícil imaginar o que poderia ser, pelo que é fácil ignorar ou rejeitar a sua proposta. Por exemplo, os linguistas descrevem frequentemente a linguagem como algo distinto de todas as outras capacidades cognitivas. Se abraçassem

a proposta de Mountcastle, talvez procurassem a semelhança entre a linguagem e a visão para melhor compreenderem a linguagem.

Para mim, esta ideia é demasiado entusiasmante para ser ignorada, e considero que as evidências empíricas a apoiam de forma esmagadora. Portanto, resta-nos um enigma fascinante: que tipo de função, ou algoritmo, pode dar origem a todos os aspetos da inteligência humana?

Até agora, descrevi uma teoria sobre como as colunas corticais aprendem modelos de objetos físicos como chávenas de café, cadeiras e smartphones. A teoria afirma que as colunas corticais criam referenciais espaciais para cada objeto observado. Recorde que um referencial é como uma grelha invisível tridimensional que envolve e se fixa a algo. Esse referencial permite que uma coluna cortical aprenda as localizações das características que definem a forma de um objeto.

Em termos mais abstratos, podemos pensar nos referenciais como uma forma de organizar qualquer tipo de conhecimento. Um referencial para uma chávena de café corresponde a um objeto físico que podemos tocar e ver. No entanto, os referenciais também podem ser usados para organizar conhecimento sobre coisas que não podemos sentir diretamente.

Pense em todas as coisas que sabe e que nunca experienciou diretamente. Por exemplo, se estudou genética, então conhece as moléculas de ADN. Consegue visualizar a sua forma de dupla hélice, sabe como codificam sequências de aminoácidos utilizando o código ATCG dos nucleótidos, e sabe como as moléculas de ADN se replicam desenrolando-se. Naturalmente, ninguém alguma vez viu ou tocou diretamente numa molécula de ADN. Não podemos, porque são demasiado pequenas. Para organizarmos o nosso conhecimento sobre as moléculas de ADN, fazemos imagens como se as pudéssemos ver e modelos como se as pudéssemos tocar. Isso permite-nos armazenar o nosso conhecimento sobre as moléculas de ADN em referenciais espaciais — da mesma forma que fazemos com o nosso conhecimento sobre chávenas de café.

Usamos este artifício para grande parte do que sabemos. Por exemplo, sabemos muito sobre fotões, e sabemos muito sobre a nossa galáxia, a Via Láctea. Mais uma vez, imaginamo-los como se

os pudéssemos ver e tocar, e por isso conseguimos organizar os factos que conhecemos sobre eles utilizando o mesmo mecanismo de referenciais que usamos para objetos físicos do quotidiano. Mas o conhecimento humano estende-se a coisas que não podem ser visualizadas. Por exemplo, temos conhecimento sobre conceitos como democracia, direitos humanos e matemática. Sabemos muitos factos sobre esses conceitos, mas não conseguimos organizar esses factos de uma forma que se assemelhe a um objeto tridimensional. Não se pode, com facilidade, criar uma imagem da democracia.

Mas deve haver alguma forma de organização para o conhecimento conceptual. Conceitos como democracia e matemática não são apenas um monte de factos. Somos capazes de raciocinar sobre eles e fazer previsões sobre o que acontecerá se agirmos de uma maneira ou de outra. A nossa capacidade de o fazer indica que o conhecimento sobre conceitos também tem de ser armazenado em referenciais. Porém, esses referenciais podem não ser facilmente equiparáveis aos referenciais que usamos para chávenas de café e outros objetos físicos. Por exemplo, é possível que os referenciais mais úteis para certos conceitos tenham mais de três dimensões. Não somos capazes de visualizar espaços com mais de três dimensões, mas, do ponto de vista matemático, funcionam da mesma forma que os espaços com três ou menos dimensões.

Todo o Conhecimento é Armazenado em Quadros de Referência

A hipótese que exploro neste capítulo é que o cérebro organiza todo o conhecimento utilizando referenciais espaciais, e que pensar é uma forma de movimento. Pensar ocorre quando ativamos localizações sucessivas dentro de referenciais.

Esta hipótese pode ser desdobrada nos seguintes componentes:

1. Os Referenciais Estão Presentes em Todo o Neocórtex

Esta premissa afirma que cada coluna do neocórtex possui células que criam referenciais. Propus que as células que o fazem são semelhantes, embora não idênticas, às células de grelha e às células de lugar encontradas em regiões mais antigas do cérebro.

2. Os Referenciais São Usados para Modelar Tudo o que Sabemos, Não Apenas Objetos Físicos

Uma coluna no neocórtex é apenas um aglomerado de neurónios. Uma coluna não "sabe" o que os seus inputs representam, nem possui qualquer conhecimento prévio sobre o que deveria aprender. Uma coluna é apenas um mecanismo constituído por neurónios que, de forma cega, tenta descobrir e modelar a estrutura daquilo que está a provocar alterações nos seus inputs.

Anteriormente, defendi que os cérebros evoluíram inicialmente referenciais para aprender a estrutura dos ambientes, de modo a podermos deslocar-nos pelo mundo. Mais tarde, os nossos cérebros evoluíram para utilizar o mesmo mecanismo na aprendizagem da estrutura dos objetos físicos, permitindo-nos reconhecê-los e manipulá-los. Agora, proponho que os nossos cérebros voltaram a evoluir no sentido de usar esse mesmo mecanismo para aprender e representar a estrutura subjacente a objetos conceptuais, como a matemática e a democracia.

3. Todo o Conhecimento Está Armazenado em Localizações Relativas a Referenciais

Os referenciais não são um componente opcional da inteligência; são a estrutura onde toda a informação é armazenada no cérebro. Cada facto que conhece está emparelhado com uma localização dentro de um referencial. Tornar-se perito numa área como a História exige a atribuição de factos históricos a localizações num referencial apropriado.

Organizar o conhecimento desta forma torna os factos passíveis de serem utilizados em ações. Recorde a analogia do mapa. Ao colocar factos sobre uma cidade num referencial em grelha, conseguimos determinar quais as ações necessárias para atingir um objetivo, como por exemplo como chegar a um determinado restaurante. A grelha uniforme do mapa torna os factos sobre a cidade operacionais. Este princípio aplica-se a todo o tipo de conhecimento.

4. Pensar é Uma Forma de Movimento

Se tudo o que sabemos está armazenado em referenciais, então, para recordarmos o conhecimento armazenado, temos de ativar as localizações apropriadas nos referenciais apropriados.

Pensar ocorre quando os neurónios invocam uma localização após outra num referencial, trazendo à mente o que foi armazenado em cada localização. A sucessão de pensamentos que experimentamos ao pensar é análoga à sucessão de sensações que temos ao tocar num objeto com um dedo, ou à sucessão de coisas que vemos quando caminhamos por uma cidade.

Os referenciais são também os meios para atingir objetivos. Tal como um mapa em papel permite descobrir como ir de onde estamos até um novo destino desejado, os referenciais no neocórtex permitem determinar os passos a dar para alcançar metas mais conceptuais, como resolver um problema de engenharia ou conseguir uma promoção no trabalho.

Embora tenhamos mencionado estas ideias sobre conhecimento conceptual em algumas das nossas publicações científicas, elas não foram o foco central, nem publicámos artigos dedicados exclusivamente a este tema. Por isso, poder-se-ia considerar que este capítulo é mais especulativo do que as partes anteriores do livro, mas não é essa a minha sensação. Apesar de existirem ainda muitos detalhes que não compreendemos, estou confiante de que o quadro geral — de que os conceitos e o

pensamento se baseiam em referenciais — resistirá à prova do tempo.

No restante deste capítulo, descreverei primeiro uma característica bem estudada do neocórtex: a sua divisão em regiões do "quê" e do "onde". Usarei esta discussão para mostrar como as colunas corticais podem desempenhar funções marcadamente diferentes através de uma simples alteração nos seus referenciais. De seguida, passo a formas de inteligência mais abstratas e conceptuais. Apresento evidências experimentais que apoiam as premissas acima e dou exemplos de como a teoria poderá relacionar-se com três tópicos: matemática, política e linguagem.

Vias do "Quê" e "Onde"

O seu cérebro possui dois sistemas visuais. Se seguir o nervo óptico desde o olho até ao neocórtex, verá que ele conduz a dois sistemas visuais paralelos, chamados a via visual do quê e a via visual do onde. A via do *quê* é um conjunto de regiões corticais que começa na parte mais posterior do cérebro e avança pelas laterais. A via do *onde* é um conjunto de regiões que também começa na parte posterior, mas move-se para cima, em direção ao topo.

As vias visuais do *quê* e do *onde* foram descobertas há mais de cinquenta anos. Anos mais tarde, os cientistas perceberam que vias paralelas semelhantes também existem para outros sentidos. Existem regiões do *quê* e do *onde* para a visão, o tato e a audição.

As vias do *quê* e do *onde* têm funções complementares. Por exemplo, se desativarmos a via visual do *onde*, então uma pessoa, ao olhar para um objeto, pode dizer o que o objeto é, mas não consegue alcançá-lo. Ela sabe que está a ver uma chávena, por exemplo, mas, de forma estranha, não consegue dizer *onde* a chávena está.

Se, por outro lado, desativarmos a via visual do *quê*, então a pessoa pode estender o braço e agarrar o objeto. Sabe *onde* ele está, mas não consegue identificar o *que é*. (Pelo menos visualmente. Quando toca no objeto com a mão, consegue identificá-lo pelo tato.)

As colunas nas regiões do *quê* e do *onde* parecem semelhantes. Têm tipos celulares, camadas e circuitos semelhantes. Então, por que razão funcionam de maneira diferente? Qual é a diferença entre uma coluna numa região do *quê* e uma coluna numa região do *onde* que leva a papéis tão distintos? Poderia ser tentado a supor que há alguma diferença no modo como os dois tipos de colunas funcionam. Talvez as colunas do *onde* tenham alguns tipos adicionais de neurónios ou ligações diferentes entre camadas. Poderia admitir que as colunas do *quê* e do *onde* se assemelham, mas argumentar que provavelmente existe alguma diferença física que ainda não foi descoberta. Se assumisse essa posição, estaria a rejeitar a proposta de Mountcastle.

Mas não é necessário abandonar a premissa de Mountcastle. Propusemos uma explicação simples para o motivo pelo qual algumas colunas são colunas do *quê* e outras são colunas do *onde*: As células de grade corticais nas colunas do *quê* atribuem referenciais aos objetos. As células de grade corticais nas colunas do *onde* atribuem referenciais ao seu corpo.

Se uma coluna visual do *onde* pudesse falar, talvez dissesse:

"Criei um referencial que está ancorado ao corpo. Usando esse referencial, olho para uma mão e sei a sua localização em relação ao corpo. Depois olho para um objeto e sei a sua localização em relação ao corpo. Com estas duas localizações, ambas no referencial do corpo, consigo calcular como mover a mão até ao objeto. Sei onde o objeto está e como alcançá-lo, mas não consigo identificá-lo. Não sei o que é aquele objeto."

Se uma coluna visual do *quê* pudesse falar, talvez dissesse:

"Criei um referencial que está ancorado a um objeto. Usando esse referencial, consigo identificar o objeto como sendo uma chávena de café. Sei o que o objeto é, mas não sei onde está."

Trabalhando em conjunto, as colunas do *quê* e do *onde* permitem-nos identificar objetos, alcançá-los e manipulá-los.

Por que razão uma coluna (coluna A) atribui referenciais a um objeto externo, enquanto outra (coluna B) os atribui ao corpo? Poderia resumir-se à origem dos inputs que chegam à coluna. Se a coluna A recebe input sensorial de um objeto, como por exemplo

sensações de um dedo a tocar numa chávena, criará automaticamente um referencial ancorado ao objeto. Se a coluna B recebe input do corpo, como neurónios que detetam os ângulos das articulações dos membros, criará automaticamente um referencial ancorado ao corpo.

De certo modo, o seu corpo é apenas mais um objeto no mundo. O neocórtex utiliza o mesmo método básico para modelar o seu corpo que usa para modelar objetos como chávenas de café. No entanto, ao contrário dos objetos externos, o seu corpo está sempre presente. Uma parte significativa do neocórtex — as regiões do *onde* — está dedicada a modelar o seu corpo e o espaço à sua volta.

A ideia de que o cérebro contém mapas do corpo não é nova. Nem tão-pouco o é a ideia de que o movimento dos membros requer referenciais centrados no corpo. O que quero salientar é que as colunas corticais, que se assemelham e funcionam de forma semelhante, podem parecer executar funções diferentes consoante os referenciais aos quais estão ancoradas. Tendo isto presente, não é um grande salto imaginar como os referenciais podem ser aplicados a conceitos.

Quadros de Referência para Conceitos

Até agora, neste livro, descrevi como o cérebro aprende modelos de coisas que têm uma forma física. Agrafadores, telemóveis, moléculas de ADN, edifícios e o seu corpo têm todos uma presença física. São todas coisas que podemos sentir diretamente ou — no caso da molécula de ADN — que conseguimos imaginar sentir.

Contudo, grande parte do que sabemos sobre o mundo não pode ser diretamente sentido e pode não ter qualquer equivalente físico. Por exemplo, não podemos estender a mão e tocar em conceitos como democracia ou números primos, e, no entanto, sabemos muito sobre essas coisas. Como é que as colunas corticais podem criar modelos de coisas que não podemos sentir?

O truque está no facto de os referenciais não precisarem de estar ancorados a algo físico. Um referencial para um conceito como

a democracia tem de ser autoconsistente, mas pode existir relativamente independente das coisas físicas do quotidiano. É semelhante à forma como podemos criar mapas de terras fictícias. Um mapa de uma terra fictícia precisa de ser autoconsistente, mas não precisa de estar localizado em nenhum ponto específico da Terra.

O segundo truque é que os referenciais para conceitos não têm de ter o mesmo número ou tipo de dimensões que os referenciais para objetos físicos como chávenas de café. As localizações dos edifícios de uma cidade descrevem-se melhor em duas dimensões. A forma de uma chávena de café descreve-se melhor em três dimensões. Mas todas as capacidades que obtemos de um referencial — como determinar a distância entre dois pontos ou calcular como nos deslocar de um local para outro — também estão presentes em referenciais com quatro ou mais dimensões.

Se tiver dificuldade em compreender como algo pode ter mais de três dimensões, considere este exemplo. Digamos que quero criar um referencial no qual possa organizar o conhecimento sobre todas as pessoas que conheço. Uma das dimensões que poderia usar seria a idade. Posso dispor os meus conhecidos ao longo dessa dimensão segundo a idade que têm. Outra métrica poderia ser o local onde vivem em relação a mim. Isso exigiria mais duas dimensões. Uma outra dimensão poderia ser a frequência com que os vejo, ou a sua altura. Já estou com cinco dimensões. Isto é apenas uma analogia; estas não seriam necessariamente as dimensões reais usadas pelo neocórtex. Mas espero que consiga ver como mais de três dimensões podem ser úteis.

É provável que as colunas do neocórtex não tenham uma noção pré-definida do tipo de referencial que devem usar. Quando uma coluna aprende um modelo de algo, parte desse processo de aprendizagem é descobrir qual é o referencial adequado — incluindo o número de dimensões.

Agora, vou rever a evidência empírica que apoia os quatro pressupostos que enumerei acima. Esta é uma área em que ainda não existe muita evidência experimental — mas já há alguma, e está a crescer.

Método dos Loci

Um truque bem conhecido para memorizar uma lista de itens, conhecido como método dos loci — ou por vezes como palácio da memória — consiste em imaginar que colocamos os itens que queremos lembrar em diferentes locais da nossa casa. Para recordar a lista, imaginamos que caminhamos pela casa, o que faz com que a memória de cada item surja, um a um. O êxito deste truque de memória mostra-nos que recordar coisas é mais fácil quando estas estão associadas a localizações dentro de um referencial familiar. Neste caso, o referencial é o mapa mental da sua casa. Repare que o ato de recordar é realizado através do movimento. Não está a mover o seu corpo fisicamente, mas sim a percorrer mentalmente a sua casa.

O método dos loci corrobora dois dos pressupostos mencionados acima: a informação é armazenada em referenciais e a recuperação da informação é uma forma de movimento. Este método é útil para memorizar rapidamente uma lista de itens, como um conjunto aleatório de substantivos. Funciona porque associa os itens a um referencial já aprendido (a sua casa) e utiliza movimentos já aprendidos (o percurso típico que faz pela casa). No entanto, na maioria das vezes, quando aprende algo, o seu cérebro cria novos referenciais. Vamos ver um exemplo disso a seguir.

Estudos com Humanos Usando fMRI

A fMRI é uma tecnologia que permite observar um cérebro vivo e ver quais as zonas que estão mais ativas. Provavelmente já viu imagens de fMRI: mostram o contorno de um cérebro com algumas partes coloridas a amarelo e vermelho, indicando onde estava a ser consumida mais energia no momento em que a imagem foi captada. A fMRI é geralmente usada com participantes humanos, porque o processo exige estar deitado perfeitamente imóvel dentro de um tubo estreito, dentro de uma máquina grande e ruidosa, enquanto se realiza uma tarefa mental específica. Muitas vezes, o sujeito está a olhar para um ecrã de computador, seguindo as instruções verbais de um investigador.

A invenção da fMRI tem sido uma mais-valia para certos tipos de investigação, mas para o tipo de investigação que fazemos não é, em geral, muito útil. A nossa pesquisa sobre a teoria do neocórtex depende de saber que neurónios individuais estão ativos em cada momento, e os neurónios ativos mudam várias vezes por segundo. Existem técnicas experimentais que fornecem este tipo de dados, mas a fMRI não tem a precisão espacial e temporal de que geralmente precisamos. A fMRI mede a atividade média de muitos neurónios e não consegue detetar atividade que dure menos de cerca de um segundo.

Por isso, ficámos surpreendidos e entusiasmados ao tomar conhecimento de uma experiência engenhosa com fMRI realizada por Christian Doeller, Caswell Barry e Neil Burgess, que demonstrou que existem células de grelha no neocórtex. Os pormenores são complexos, mas os investigadores perceberam que as células de grelha podiam exibir uma assinatura detetável através da fMRI. Primeiro, tinham de verificar se a técnica funcionava, por isso focaram-se no córtex entorrinal, onde se sabe que existem células de grelha. Pediram aos participantes humanos que realizassem uma tarefa de navegação, movendo-se num mundo virtual num ecrã de computador e, usando fMRI, conseguiram detetar a presença de atividade típica de células de grelha enquanto os sujeitos realizavam a tarefa. Depois voltaram a sua atenção para o neocórtex. Aplicaram a técnica de fMRI para observar áreas frontais do neocórtex enquanto o sujeito executava a mesma tarefa de navegação. Encontraram a mesma assinatura, sugerindo fortemente que também existem células de grelha em pelo menos algumas regiões do neocórtex.

Outra equipa de cientistas, Alexandra Constantinescu, Jill O'Reilly e Timothy Behrens, utilizou a nova técnica de fMRI para uma tarefa diferente. Os sujeitos viam imagens de pássaros. Os pássaros variavam no comprimento do pescoço e das pernas. Os participantes eram convidados a realizar várias tarefas de imaginação mental relacionadas com os pássaros, como, por exemplo, imaginar um novo pássaro que combinasse características de dois pássaros anteriormente observados. As experiências mostraram não só que existem células de grelha em áreas frontais do neocórtex, mas também forneceram indícios de que o neocórtex

armazenava as imagens dos pássaros num referencial com estrutura de mapa — uma dimensão representava o comprimento do pescoço e outra o comprimento das pernas. A equipa de investigação demonstrou ainda que, quando os sujeitos pensavam nos pássaros, estavam a "mover-se" mentalmente através do mapa dos pássaros da mesma forma que se pode percorrer mentalmente o mapa da própria casa. Mais uma vez, os detalhes da experiência são complexos, mas os dados da fMRI sugerem que essa parte do neocórtex usava neurónios do tipo células de grelha para aprender sobre os pássaros. Os sujeitos que participaram nesta experiência não faziam ideia de que isto estava a acontecer, mas os dados das imagens eram inequívocos.

O método dos loci utiliza um mapa previamente aprendido — o mapa da sua casa — para armazenar itens que poderão ser recordados mais tarde. No exemplo dos pássaros, o neocórtex criou um novo mapa, um mapa adaptado à tarefa de memorizar pássaros com diferentes comprimentos de pescoço e pernas. Em ambos os exemplos, o processo de armazenar itens num referencial e recordá-los através de "movimento" é o mesmo.

Se todo o conhecimento é armazenado desta forma, então aquilo a que normalmente chamamos pensamento é, na verdade, um movimento através de um espaço, através de um referencial. O seu pensamento atual — aquilo que está na sua mente em determinado momento — é determinado pela localização atual no referencial. À medida que essa localização muda, os itens armazenados em cada localização são recordados um a um. Os nossos pensamentos estão em constante mudança, mas não são aleatórios. Aquilo em que pensamos a seguir depende da direção em que nos movemos mentalmente no referencial, da mesma forma que o que vemos a seguir numa cidade depende da direção para onde nos deslocamos a partir da nossa localização atual.

O referencial necessário para aprender o que é uma chávena de café é talvez evidente: trata-se do espaço tridimensional em torno da chávena. O referencial aprendido na experiência de fMRI sobre os pássaros é talvez um pouco menos óbvio. Mas o referencial dos pássaros continua a estar relacionado com atributos físicos dos pássaros, como as pernas e os pescoços. Mas que tipo de referencial

deve o cérebro usar para conceitos como economia ou ecologia? Pode haver múltiplos referenciais possíveis, embora alguns sejam mais eficazes do que outros.

Esta é uma das razões pelas quais aprender conhecimento conceptual pode ser difícil. Se eu lhe der dez acontecimentos históricos relacionados com a democracia, como deverá organizá-los? Um professor poderá apresentar os eventos dispostos numa linha temporal. Uma linha temporal é um referencial unidimensional. É útil para avaliar a ordem temporal dos acontecimentos e perceber que eventos poderão estar causalmente ligados pela proximidade temporal. Outro professor poderá organizar os mesmos eventos históricos geograficamente num mapa-mundo. Um referencial geográfico sugere diferentes formas de pensar sobre os mesmos acontecimentos, como, por exemplo, que eventos podem estar ligados causalmente por proximidade espacial, ou por estarem próximos de oceanos, desertos ou montanhas. Linhas temporais e geografia são ambas formas válidas de organizar acontecimentos históricos, e cada uma conduz a diferentes maneiras de pensar sobre a história. Podem conduzir a conclusões e previsões distintas. A melhor estrutura para aprender sobre democracia poderá exigir um mapa totalmente novo, um mapa com múltiplas dimensões abstratas que correspondam, por exemplo, a justiça ou direitos. Não estou a sugerir que "justiça" ou "direitos" sejam dimensões reais usadas pelo cérebro. O meu ponto é que tornar-se perito numa área de estudo requer a descoberta de uma boa estrutura para representar os dados e factos associados. Pode não haver um referencial "correto", e duas pessoas diferentes poderão organizar os factos de forma distinta. Descobrir um referencial útil é a parte mais difícil da aprendizagem, mesmo que, na maioria das vezes, não tenhamos consciência disso. Irei ilustrar esta ideia com os três exemplos que referi anteriormente: matemática, política e linguagem.

Matemática

Suponhamos que é matemático e quer provar a conjetura OMG (OMG não é uma conjetura real). Uma conjetura é uma afirmação matemática que se acredita ser verdadeira, mas que ainda não foi

provada. Para provar uma conjetura, começa com algo que se sabe ser verdadeiro. Depois aplica uma série de operações matemáticas. Se, através desse processo, chegar a uma afirmação que é a conjetura, então conseguiu prová-la. Normalmente, haverá uma série de resultados intermédios. Por exemplo, começando por A, prova-se B. De B, prova-se C. E finalmente, de C, prova-se OMG. Digamos que A, B, C e a OMG final são equações. Para passar de equação em equação, tem de realizar uma ou mais operações matemáticas.

Agora suponhamos que as várias equações estão representadas no seu neocórtex num referencial. Operações matemáticas, como multiplicar ou dividir, são movimentos que lhe levam a diferentes locais neste referencial. Executar uma série de operações move-o para uma nova localização, uma nova equação. Se conseguir determinar um conjunto de operações — movimentos pelo espaço das equações — que lhe leve de A até à OMG, então conseguiu provar a OMG.

Resolver problemas complexos, como uma conjetura matemática, requer muito treino. Ao aprender um novo campo, o seu cérebro não está apenas a armazenar factos. Para a matemática, o cérebro deve descobrir referenciais úteis onde guardar as equações e os números, e deve aprender como os comportamentos matemáticos, como operações e transformações, se traduzem em movimentos para novas localizações dentro desses referenciais.

Para um matemático, as equações são objetos familiares, tal como você e eu vemos um smartphone ou uma bicicleta. Quando os matemáticos veem uma nova equação, reconhecem-na como semelhante a equações anteriores com as quais já trabalharam, e isso sugere imediatamente como podem manipular a nova equação para obter certos resultados. É o mesmo processo que seguimos se vemos um novo smartphone. Reconhecemos que o telefone é parecido com outros que já usamos e isso sugere como o podemos manipular para alcançar um resultado desejado.

No entanto, se não está treinado em matemática, as equações e outras notações matemáticas vão parecer-lhe rabiscos sem sentido. Poderá até reconhecer uma equação que já viu antes, mas

sem um referencial não terá ideia de como a manipular para resolver um problema. Pode sentir-se perdido no espaço da matemática, da mesma forma que pode sentir-se perdido numa floresta sem mapa.

Matemáticos a manipular equações, exploradores a viajar pela floresta e dedos a tocar chávenas de café precisam todos de referenciais do tipo mapa para saber onde estão e que movimentos precisam fazer para chegar onde querem. O mesmo algoritmo básico está na base destas e de inúmeras outras atividades que realizamos.

Política

O exemplo matemático acima é completamente abstrato, mas o processo é o mesmo para qualquer problema que não seja manifestamente físico. Por exemplo, imagine-se que um político quer fazer aprovar uma nova lei. Ele tem um primeiro esboço da lei escrito, mas existem vários passos necessários para atingir o objetivo final da aprovação. Pelo caminho há obstáculos políticos, por isso o político pensa em todas as diferentes ações que poderá tomar. Um político experiente sabe o que provavelmente acontecerá se der uma conferência de imprensa, ou obrigar a um referendo, ou escrever um documento político, ou oferecer trocar apoio por outro projeto de lei. Um político habilidoso aprendeu um referencial para a política. Parte desse referencial é como as ações políticas mudam as localizações dentro do referencial, e o político imagina o que acontecerá se fizer estas coisas. O seu objetivo é encontrar uma série de ações que o levem ao resultado desejado: aprovar a nova lei.

Um político e um matemático não têm consciência de que estão a usar referenciais para organizar o seu conhecimento, tal como você e eu não temos consciência de que usamos referenciais para compreender smartphones e grampeadores. Não andamos por aí a perguntar, "Alguém pode sugerir um referencial para organizar estes factos?" O que dizemos é: "Preciso de ajuda. Não entendo como resolver este problema." Ou "Estou confuso. Podes mostrar-me como usar esta coisa?" Ou "Estou perdido. Podes mostrar-me como chegar à cantina?" Estas são as perguntas que fazemos

quando não conseguimos atribuir um referencial aos factos que temos à nossa frente.

Língua

A linguagem é, indiscutivelmente, a capacidade cognitiva mais importante que distingue os humanos de todos os outros animais. Sem a capacidade de partilhar conhecimento e experiências através da linguagem, grande parte da sociedade moderna não seria possível.

Embora tenham sido escritos muitos volumes sobre linguagem, desconheço tentativas de explicar como a linguagem é criada pelos circuitos neuronais observados no cérebro. Os linguistas normalmente não se aventuram na neurociência e, embora alguns neurocientistas estudem regiões cerebrais relacionadas com a linguagem, não têm sido capazes de propor teorias detalhadas sobre como o cérebro cria e compreende a linguagem.

Existe um debate contínuo sobre se a linguagem é fundamentalmente diferente das outras capacidades cognitivas. Os linguistas tendem a pensar que sim. Eles descrevem a linguagem como uma capacidade única, diferente de tudo o que fazemos. Se isso fosse verdade, as partes do cérebro que criam e compreendem a linguagem deveriam ser diferentes. Aqui, a neurociência é ambígua.

Existem duas regiões de tamanho moderado no neocortex que se dizem ser responsáveis pela linguagem. A área de Wernicke é considerada responsável pela compreensão da linguagem, e a área de Broca é considerada responsável pela produção da linguagem. Isto é uma simplificação. Em primeiro lugar, há desacordo quanto à localização exata e extensão destas regiões. Em segundo lugar, as funções das áreas de Wernicke e Broca não se distinguem claramente entre compreensão e produção; elas sobrepõem-se um pouco. Finalmente, e isto deveria ser óbvio, a linguagem não pode ser isolada a duas pequenas regiões do neocórtex. Usamos linguagem falada, escrita e gestual. As áreas de Wernicke e Broca não recebem input diretamente dos sensores, por isso a

compreensão da linguagem depende das regiões auditivas e visuais, e a produção da linguagem depende de diferentes capacidades motoras. São necessárias grandes áreas do neocórtex para criar e compreender a linguagem. As áreas de Wernicke e Broca desempenham um papel-chave, mas é errado pensar que criam a linguagem isoladamente.

Um facto surpreendente sobre a linguagem, que sugere que a linguagem pode ser diferente das outras funções cognitivas, é que as áreas de Broca e Wernicke existem apenas no lado esquerdo do cérebro. As áreas equivalentes no lado direito estão apenas marginalmente implicadas na linguagem. Quase tudo o resto que o neocórtex faz ocorre em ambos os lados do cérebro. A assimetria única da linguagem sugere que há algo diferente nas áreas de Broca e Wernicke.

Por que a linguagem ocorre apenas no lado esquerdo do cérebro pode ter uma explicação simples. Uma proposta é que a linguagem exige um processamento rápido, e os neurónios na maior parte do neocórtex são demasiado lentos para processar linguagem. Os neurónios nas áreas de Wernicke e Broca são conhecidos por ter uma camada extra de isolamento (chamada mielina) que lhes permite funcionar mais rapidamente e acompanhar as exigências da linguagem. Existem outras diferenças notórias em relação ao resto do neocórtex. Por exemplo, foi reportado que o número e a densidade de sinapses são maiores nas regiões da linguagem comparadas com as suas equivalentes no lado direito do cérebro. Mas ter mais sinapses não significa que as áreas da linguagem desempenhem uma função diferente; pode simplesmente significar que estas áreas aprenderam mais coisas.

Embora existam algumas diferenças, a anatomia das áreas de Wernicke e Broca é, mais uma vez, semelhante a outras áreas do neocórtex. Os factos que temos hoje sugerem que, embora estas áreas da linguagem sejam um pouco diferentes, talvez de forma subtil, a estrutura geral das camadas, da conetividade e dos tipos celulares é semelhante ao resto do neocórtex. Portanto, a maioria dos mecanismos subjacentes à linguagem é provavelmente partilhada com outras partes da cognição e da perceção. Esta deve ser a nossa hipótese de trabalho até que se prove o contrário.

Assim, podemos perguntar: como é que as capacidades de modelação de uma coluna cortical, incluindo os referenciais, podem fornecer uma base para a linguagem?

De acordo com os linguistas, um dos atributos definidores da linguagem é a sua estrutura aninhada. Por exemplo, as frases são compostas por expressões, as expressões são compostas por palavras, e as palavras são compostas por letras. A recursividade, a capacidade de aplicar repetidamente uma regra, é outro atributo definidor. A recursividade permite construir frases com complexidade quase ilimitada. Por exemplo, a frase simples "Tom pediu mais chá" pode ser alargada para "Tom, que trabalha na oficina automóvel, pediu mais chá", que pode ser ainda alargada para "Tom, que trabalha na oficina automóvel, aquela ao lado da loja de segunda mão, pediu mais chá." A definição exata de recursividade no contexto da linguagem ainda não reúne consenso, mas a ideia geral é fácil de entender. As frases podem ser compostas por expressões, que podem ser compostas por outras expressões, e assim sucessivamente. Há muito que se argumenta que a estrutura aninhada e a recursividade são atributos chave da linguagem.

No entanto, a estrutura aninhada e recursiva não é exclusiva da linguagem. De facto, tudo no mundo é composto desta forma. Pegue no meu copo de café com o logótipo da Numenta impresso na lateral. O copo tem uma estrutura aninhada: consiste num cilindro, uma pega e um logótipo. O logótipo consiste numa imagem e numa palavra. A imagem é composta por círculos e linhas, enquanto a palavra "Numenta" é composta por sílabas, e as sílabas são elas próprias feitas de letras. Os objetos podem também ter uma estrutura recursiva. Por exemplo, imagine que o logótipo da Numenta incluía uma imagem de um copo de café, no qual estava impresso um logótipo da Numenta, que por sua vez tinha uma imagem de um copo de café, etc.

No início da nossa investigação, percebemos que cada coluna cortical tinha de ser capaz de aprender estrutura aninhada e recursiva. Esta era uma condição necessária para aprender a estrutura de coisas físicas como copos de café e para aprender a estrutura de coisas conceptuais como matemática e linguagem.

Qualquer teoria que concebêssemos teria de explicar como as colunas realizam isto.

Imagine que, em algum momento no passado, aprendeu como é um copo de café, e também aprendeu como é o logótipo da Numenta. Mas nunca tinha visto o logótipo num copo de café. Agora, mostro-lhe um novo copo de café com o logótipo na lateral. Consegue aprender rapidamente o novo objeto combinado, normalmente com apenas um ou dois olhares. Note que não precisa de reaprender o logótipo nem o copo. Tudo o que sabemos sobre copos e sobre o logótipo é imediatamente incluído como parte do novo objeto.

Como é que isto acontece? Dentro de uma coluna cortical, o copo de café aprendido anteriormente está definido por um referencial. O logótipo aprendido anteriormente está também definido por um referencial. Para aprender o copo de café com o logótipo, a coluna cria um novo referencial, no qual armazena duas coisas: uma ligação para o referencial do copo previamente aprendido e uma ligação para o referencial do logótipo previamente aprendido. O cérebro consegue fazer isto rapidamente, com apenas algumas sinapses adicionais. Isto é um pouco como usar hiperligações num documento de texto. Imagine que escrevi um pequeno ensaio sobre Abraham Lincoln e menciono que ele proferiu um famoso discurso chamado Gettysburg Address. Ao transformar as palavras "Gettysburg Address" numa ligação para o discurso completo, posso incluir todos os detalhes do discurso como parte do meu ensaio sem precisar de o reescrever.

Anteriormente disse que as colunas corticais armazenam características em localizações nos referenciais. A palavra "característica" é um pouco vaga. Agora serei mais preciso. As colunas corticais criam referenciais para cada objeto que conhecem. Os referenciais são então preenchidos com ligações para outros referenciais. O cérebro modela o mundo usando referenciais que são preenchidos com referenciais; são referenciais a todo o comprimento. No nosso artigo de 2019, "*Frameworks*", propusemos como os neurónios poderão fazer isto.

Ainda temos um longo caminho a percorrer para entender completamente tudo o que o neocórtex faz. Mas a ideia de que cada

coluna modela objetos usando referenciais é, até onde sabemos, consistente com as necessidades da linguagem. Talvez mais à frente venhamos a encontrar necessidade de alguns circuitos especiais para a linguagem. Mas, para já, não é o caso.

Especialização

Até agora, apresentei quatro usos para os referenciais: um no cérebro antigo e três no neocórtex. Os referenciais no cérebro antigo aprendem mapas dos ambientes. Os referenciais nas colunas do neocórtex da via "o quê" aprendem mapas de objetos físicos. Os referenciais nas colunas da via "onde" do neocórtex aprendem mapas do espaço em redor do nosso corpo. E, finalmente, os referenciais nas colunas não sensoriais do neocórtex aprendem mapas de conceitos.

Ser especialista em qualquer domínio requer ter um bom referencial, um bom mapa. Duas pessoas a observar o mesmo objeto físico provavelmente acabarão por ter mapas semelhantes. Por exemplo, é difícil imaginar como os cérebros de duas pessoas a observar a mesma cadeira organizariam as suas características de forma diferente. Mas, quando pensamos em conceitos, duas pessoas que partem dos mesmos factos podem acabar com referenciais diferentes. Recorde-se o exemplo de uma lista de factos históricos. Uma pessoa pode organizar os factos numa linha temporal, e outra pode organizá-los num mapa geográfico. Os mesmos factos podem levar a modelos diferentes e a visões de mundo distintas.

Ser especialista é, em grande parte, encontrar um bom referencial para organizar factos e observações. Albert Einstein partiu dos mesmos factos que os seus contemporâneos. No entanto, ele encontrou uma melhor forma de os organizar, um melhor referencial, que lhe permitiu ver analogias e fazer previsões surpreendentes. O mais fascinante nas descobertas de Einstein relacionadas com a relatividade especial é que os referenciais que usou para as fazer eram objetos do quotidiano. Pensou em comboios, pessoas e lanternas. Começou pelas observações empíricas de cientistas, como a velocidade absoluta da luz, e usou referenciais do dia a dia para deduzir as equações da relatividade

especial. Por causa disso, quase qualquer pessoa pode seguir a sua lógica e compreender como fez as suas descobertas. Em contraste, a teoria geral da relatividade de Einstein exigia referenciais baseados em conceitos matemáticos chamados equações de campo, que não se relacionam facilmente com objetos do quotidiano. Einstein achou isto muito mais difícil de compreender, como praticamente toda a gente.

Em 1978, quando Vernon Mountcastle propôs que existia um algoritmo comum subjacente a toda a perceção e cognição, era difícil imaginar que algoritmo poderia ser suficientemente poderoso e geral para cumprir o requisito. Era difícil imaginar um único processo que pudesse explicar tudo aquilo a que chamamos inteligência, desde a perceção sensorial básica até às formas mais elevadas e admiradas de capacidade intelectual. Agora está claro para mim que o algoritmo cortical comum se baseia em referenciais. Os referenciais fornecem a base para aprender a estrutura do mundo, onde as coisas estão, e como se movem e mudam. Os referenciais conseguem fazer isto não só para os objetos físicos que podemos sentir diretamente, mas também para objetos que não podemos ver nem tocar, e até para conceitos que não têm forma física.

O seu cérebro tem 150 000 colunas corticais. Cada coluna é uma máquina de aprendizagem. Cada coluna aprende um modelo preditivo das suas entradas observando como estas mudam ao longo do tempo. As colunas não sabem o que estão a aprender; não sabem o que os seus modelos representam. Todo o empreendimento e os modelos resultantes são construídos em referenciais. O referencial correto para entender como o cérebro funciona são os referenciais.

CAPÍTULO 7

A Teoria dos Mil Cérebros da Inteligência

Desde o seu início, o objetivo da Numenta foi desenvolver uma teoria abrangente sobre o funcionamento do neocórtex. Os neurocientistas publicavam milhares de artigos por ano, cobrindo todos os detalhes do cérebro, mas havia uma falta de teorias sistémicas que articulassem esses detalhes de forma coerente. Decidimos começar por compreender uma única coluna cortical. Sabíamos que as colunas corticais eram fisicamente complexas e, por isso, deviam realizar algo igualmente complexo. Não fazia sentido perguntar por que é que as colunas estão ligadas entre si daquela forma confusa e algo hierárquica que mostrei no Capítulo 2, se não soubéssemos o que faz uma única coluna. Isso seria como tentar perceber como funcionam as sociedades sem saber nada sobre as pessoas.

Agora sabemos muito sobre o que fazem as colunas corticais. Sabemos que cada coluna é um sistema sensório-motor. Sabemos que cada coluna pode aprender modelos de centenas de objetos, e que esses modelos se baseiam em referenciais. Uma vez compreendido que as colunas fazem estas coisas, tornou-se evidente que o neocórtex, como um todo, funciona de maneira diferente do que anteriormente se pensava. Chamamos a esta nova perspetiva a Teoria dos Mil Cérebros da Inteligência. Antes de explicar o que é a Teoria dos Mil Cérebros, será útil compreender o que ela vem substituir.

A Visão Existente do Neocórtex

Hoje em dia, a forma mais comum de se pensar sobre o neocórtex é como um organograma. A informação proveniente dos sentidos é processada passo a passo à medida que passa de uma região do neocórtex para a seguinte. Os cientistas referem-se a isto

como uma hierarquia de detetores de características. É mais frequentemente descrita em termos de visão, e funciona da seguinte forma: cada célula da retina deteta a presença de luz numa pequena parte de uma imagem. As células da retina projetam então para o neocórtex. A primeira região do neocórtex que recebe esta informação é chamada de região V1. Cada neurónio da região V1 recebe input de apenas uma pequena parte da retina. É como se estivessem a olhar para o mundo através de uma palhinha.

Estes factos sugerem que as colunas na região V1 não conseguem reconhecer objetos completos. Portanto, o papel de V1 está limitado à deteção de pequenas características visuais, como linhas ou contornos, numa parte localizada da imagem. Em seguida, os neurónios de V1 transmitem estas características para outras regiões do neocórtex. A próxima região visual, chamada V2, combina as características simples provenientes de V1 em características mais complexas, como cantos ou arcos. Este processo repete-se mais algumas vezes, em mais algumas regiões, até que os neurónios respondam a objetos completos. Presume-se que um processo semelhante — do simples ao complexo até chegar a objetos inteiros — ocorre também com o tato e a audição. Esta visão do neocórtex como uma hierarquia de detetores de características tem sido a teoria dominante há cinquenta anos.

O maior problema com esta teoria é que ela trata a visão como um processo estático, como tirar uma fotografia. Mas a visão não é assim. Cerca de três vezes por segundo, os nossos olhos fazem movimentos rápidos, chamados sacádicos. Os inputs dos olhos para o cérebro mudam completamente com cada sacada. Os inputs visuais também mudam quando andamos para a frente ou viramos a cabeça para a esquerda ou para a direita. A teoria da hierarquia de características ignora estas mudanças. Trata a visão como se o objetivo fosse tirar uma fotografia de cada vez e rotulá-la. Mas mesmo uma observação casual dirá que a visão é um processo interativo, dependente do movimento. Por exemplo, para aprender como é um novo objeto, seguramo-lo na mão, rodando-o de um lado para o outro, para ver como é de diferentes ângulos. Só através do movimento podemos aprender um modelo do objeto.

Uma razão pela qual muitas pessoas ignoraram o aspeto dinâmico da visão é que, por vezes, conseguimos reconhecer uma imagem sem mover os olhos, como uma imagem brevemente mostrada num ecrã — mas isso é uma exceção, não a regra. A visão normal é um processo sensório-motor ativo, não um processo estático.

O papel essencial do movimento é ainda mais evidente no tato e na audição. Se alguém coloca um objeto na sua mão aberta, não o conseguirá identificar sem mover os dedos. Da mesma forma, a audição é sempre dinâmica. Não só os objetos auditivos, como palavras faladas, são definidos por sons que mudam ao longo do tempo, como também, ao ouvirmos, movemos a cabeça para modificar ativamente o que escutamos. Não é claro como a teoria da hierarquia de características se aplica sequer ao tato ou à audição. Com a visão, pelo menos, podemos imaginar que o cérebro está a processar uma imagem semelhante a uma fotografia, mas com o tato e a audição não há nada equivalente.

Existem numerosas outras observações que sugerem que a teoria da hierarquia de características precisa de ser modificada. Eis algumas, todas relacionadas com a visão:

- As primeiras e segundas regiões visuais, V1 e V2, estão entre as maiores do neocórtex humano. São substancialmente maiores em área do que outras regiões visuais, onde supostamente são reconhecidos os objetos completos. Por que exigiria a deteção de pequenas características, que são limitadas em número, uma fração maior do cérebro do que o reconhecimento de objetos completos, dos quais há muitos? Em alguns mamíferos, como o rato, este desequilíbrio é ainda mais acentuado. A região V1 no rato ocupa uma grande porção de todo o neocórtex do animal. Outras regiões visuais no rato são minúsculas em comparação. É como se quase toda a visão do rato ocorresse na região V1.
- Os neurónios detetores de características em V1 foram descobertos quando investigadores projetaram imagens diante dos olhos de animais anestesiados, ao

mesmo tempo que registavam a atividade dos neurónios em V1. Encontraram neurónios que se tornavam ativos perante características simples, como uma aresta, numa pequena parte da imagem. Como os neurónios apenas respondiam a características simples numa área reduzida, assumiu-se que os objetos completos teriam de ser reconhecidos noutra parte. Isto levou ao modelo hierárquico de características. Mas, nestas experiências, a maioria dos neurónios em V1 não respondia a nada de óbvio — podiam emitir um impulso ocasionalmente, ou podiam disparar continuamente durante algum tempo e depois parar. A maioria dos neurónios não podia ser explicada pela teoria da hierarquia de características, e por isso foram largamente ignorados. No entanto, todos esses neurónios não contabilizados em V1 devem estar a fazer algo importante que não é a deteção de características.

- Quando os olhos executam um movimento sacádico de um ponto de fixação para outro, alguns dos neurónios nas regiões V1 e V2 fazem algo notável. Parecem saber o que vão ver antes de os olhos pararem de se mover. Estes neurónios tornam-se ativos como se pudessem ver o novo input, mesmo antes de este ter chegado. Os cientistas que descobriram isto ficaram surpreendidos. Isso implicava que os neurónios nas regiões V1 e V2 tinham acesso ao conhecimento sobre o objeto inteiro que estava a ser visto, e não apenas sobre uma pequena parte dele.
- Existem mais fotorrecetores no centro da retina do que na periferia. Se pensarmos no olho como uma câmara, então trata-se de uma com uma lente "olho de peixe" extremamente acentuada. Existem também partes da retina que não têm fotorrecetores, por exemplo, o ponto cego, onde o nervo ótico sai do olho e onde os vasos sanguíneos cruzam a retina.

Consequentemente, o input para o neocórtex não é como uma fotografia. É uma colagem altamente distorcida e incompleta de fragmentos de imagem. No entanto, não temos consciência destas distorções e omissões; a nossa perceção do mundo é uniforme e completa. A teoria da hierarquia de características não consegue explicar como isto acontece. Este problema é conhecido como o problema da integração ou problema da fusão sensorial. Mais genericamente, o problema da integração pergunta como é que inputs provenientes de diferentes sentidos, que estão espalhados por todo o neocórtex com todo o tipo de distorções, são combinados numa perceção única e não distorcida, como a que todos experienciamos.

- Tal como apontei no Capítulo 1, embora algumas das ligações entre regiões do neocórtex pareçam hierárquicas, como num fluxograma passo a passo, a maioria não o é. Por exemplo, existem ligações entre regiões visuais de baixo nível e regiões táteis também de baixo nível. Estas ligações não fazem sentido dentro da teoria da hierarquia de características.
- Embora a teoria da hierarquia de características possa explicar como o neocórtex reconhece uma imagem, não fornece qualquer pista sobre como aprendemos a estrutura tridimensional dos objetos, como os objetos são compostos por outros objetos, e como os objetos mudam e se comportam ao longo do tempo. Não explica como conseguimos imaginar como será um objeto ao ser rodado ou distorcido.

Com todas estas inconsistências e limitações, talvez se esteja a perguntar por que razão a teoria da hierarquia de características continua a ser amplamente aceite. Há várias razões. Em primeiro lugar, adapta-se a muitos dados, especialmente aos que foram recolhidos há muito tempo. Em segundo lugar, os problemas com a teoria acumularam-se lentamente ao longo do tempo, tornando fácil descartar cada novo problema como algo menor. Em terceiro lugar, é a melhor teoria que temos e, na ausência de algo que a substitua,

mantemo-nos fiéis a ela. Por fim, como argumentarei em breve, não está completamente errada — apenas precisa de uma grande atualização.

A Nova Visão do Neocórtex

A nossa proposta de quadros de referência nas colunas corticais sugere uma forma diferente de pensar sobre o funcionamento do neocórtex. Diz que todas as colunas corticais, mesmo nas regiões sensoriais de baixo nível, são capazes de aprender e reconhecer objetos completos. Uma coluna que capta apenas uma pequena parte de um objeto pode aprender um modelo do objeto inteiro ao integrar os seus inputs ao longo do tempo, da mesma forma que você e eu aprendemos uma nova cidade visitando um local após outro. Assim, uma hierarquia de regiões corticais não é estritamente necessária para aprender modelos de objetos. A nossa teoria explica como um rato, com um sistema visual praticamente monocamadal, pode ver e reconhecer objetos no mundo.

O neocórtex possui muitos modelos de um determinado objeto. Os modelos estão em diferentes colunas. Não são idênticos, mas complementares. Por exemplo, uma coluna que recebe input tátil de uma ponta de dedo pode aprender um modelo de um telemóvel que inclua a sua forma, as texturas das suas superfícies e como os seus botões se movem quando pressionados. Uma coluna que recebe input visual da retina pode aprender um modelo do mesmo telemóvel que também inclua a sua forma, mas, ao contrário da coluna tátil, o seu modelo pode incluir a cor das diferentes partes do telefone e como os ícones visuais no ecrã mudam à medida que o utiliza. Uma coluna visual não consegue aprender a sensação do clique do botão de ligar/desligar, e uma coluna tátil não consegue aprender como os ícones mudam no ecrã.

Nenhuma coluna cortical individual consegue aprender um modelo de todos os objetos do mundo. Isso seria impossível. Em primeiro lugar, existe um limite físico para quantos objetos uma única coluna pode aprender. Ainda não sabemos qual é esse limite, mas as nossas simulações sugerem que uma única coluna pode aprender centenas de objetos complexos. Isto é muito menos do

que o número de coisas que você conhece. Além disso, o que uma coluna aprende é limitado pelos seus inputs. Por exemplo, uma coluna tátil não pode aprender modelos de nuvens, e uma coluna visual não pode aprender melodias.

Mesmo dentro de um único sentido, como a visão, as colunas recebem diferentes tipos de input e, por isso, aprenderão diferentes tipos de modelos. Por exemplo, há colunas visuais que recebem input de cor e outras que recebem input a preto e branco. Noutro exemplo, as colunas nas regiões V1 e V2 recebem ambas input da retina. Uma coluna na região V1 recebe input de uma área muito pequena da retina, como se estivesse a ver o mundo através de uma palhinha estreita. Uma coluna em V2 recebe input de uma área maior da retina, como se estivesse a ver o mundo através de uma palhinha mais larga, mas com a imagem mais desfocada.

Agora imagine que está a olhar para texto no mais pequeno tipo de letra que consegue ler. A nossa teoria sugere que apenas colunas na região V1 conseguem reconhecer letras e palavras nesse tipo de letra minúsculo. A imagem vista por V2 é demasiado desfocada. À medida que aumentamos o tamanho da letra, então tanto V2 como V1 conseguem reconhecer o texto. Se o tipo de letra aumentar ainda mais, então torna-se mais difícil para V1 reconhecer o texto, mas V2 continua a conseguir fazê-lo. Assim, colunas nas regiões V1 e V2 podem ambas aprender modelos de objetos, como letras e palavras, mas os modelos diferem em escala.

Onde Está Armazenado o Conhecimento no Cérebro?

O conhecimento no cérebro está distribuído. Nada do que sabemos está armazenado num único local, como uma única célula ou uma única coluna. Também não está armazenado em todo o lado, como num holograma. O conhecimento sobre algo está distribuído por milhares de colunas, mas estas constituem apenas uma pequena fração de todas as colunas.

Voltemos a considerar o nosso exemplo da chávena de café. Onde está armazenado, no cérebro, o conhecimento sobre a chávena de café? Existem muitas colunas corticais nas regiões

visuais que recebem input da retina. Cada coluna que observa uma parte da chávena aprende um modelo da chávena e tenta reconhecê-la. Do mesmo modo, se agarrar a chávena com as mãos, dezenas ou centenas de modelos nas regiões táteis do neocórtex tornam-se ativos. Não existe um modelo único de chávenas de café. O que sabe sobre chávenas de café existe em milhares de modelos, em milhares de colunas — mas, ainda assim, apenas numa fração das colunas do neocórtex. É por isso que chamamos a isto a Teoria dos Mil Cérebros: o conhecimento de qualquer item em particular está distribuído por milhares de modelos complementares.

Eis uma analogia. Imaginemos uma cidade com cem mil habitantes. A cidade tem uma rede de canos, bombas, tanques e filtros para fornecer água potável a cada casa. Este sistema de águas precisa de manutenção para se manter em bom funcionamento. Onde reside o conhecimento sobre como manter o sistema? Seria imprudente que apenas uma pessoa soubesse fazê-lo, e seria pouco prático que todos os cidadãos o soubessem. A solução é distribuir o conhecimento por muitas pessoas — mas não por demasiadas. Neste caso, digamos que o departamento das águas tem cinquenta funcionários. Continuando com esta analogia, digamos que o sistema de águas tem cem componentes — isto é, cem bombas, válvulas, tanques, etc. — e que cada um dos cinquenta trabalhadores do departamento sabe como manter e reparar um conjunto diferente, mas sobreposto, de vinte desses componentes.

Então, onde está armazenado o conhecimento sobre o sistema de águas? Cada um dos cem componentes é conhecido por cerca de dez pessoas diferentes. Se metade dos funcionários faltasse ao trabalho num determinado dia, seria ainda assim altamente provável que houvesse pelo menos cinco pessoas capazes de reparar qualquer componente. Cada funcionário consegue manter e reparar 20 por cento do sistema por si só, sem supervisão. O conhecimento de como manter e reparar o sistema de águas está distribuído por uma pequena parte da população, e esse conhecimento é robusto face a uma perda significativa de pessoal.

Repare que o departamento das águas pode ter alguma hierarquia de controlo, mas seria imprudente impedir qualquer

autonomia ou atribuir um pedaço de conhecimento apenas a uma ou duas pessoas. Sistemas complexos funcionam melhor quando o conhecimento e as ações estão distribuídos por muitos elementos — mas não por demasiados.

Tudo no cérebro funciona assim. Por exemplo, um neurónio nunca depende de uma única sinapse. Em vez disso, pode usar trinta sinapses para reconhecer um padrão. Mesmo que dez dessas sinapses falhem, o neurónio continuará a reconhecer o padrão. Uma rede de neurónios nunca depende de uma única célula. Nas redes simuladas que criámos, mesmo a perda de 30 por cento dos neurónios tem geralmente apenas um efeito marginal no desempenho da rede. Do mesmo modo, o neocórtex não depende de uma única coluna cortical. O cérebro continua a funcionar mesmo que um AVC ou um trauma elimine milhares de colunas.

Portanto, não devemos ficar surpreendidos pelo facto de o cérebro não depender de um único modelo de nada. O nosso conhecimento sobre algo está distribuído por milhares de colunas corticais. As colunas não são redundantes, nem são cópias exatas umas das outras. O mais importante é que cada coluna é um sistema sensório-motor completo, tal como cada trabalhador do departamento das águas é capaz de reparar, de forma autónoma, uma parte da infraestrutura hídrica.

A Solução para o Problema da Ligação

Por que temos uma perceção singular se temos milhares de modelos? Quando seguramos e olhamos para uma chávena de café, por que é que a chávena nos parece uma coisa única, e não milhares de coisas? Se pousarmos a chávena numa mesa e ela fizer um som, como é que esse som se une à imagem e à sensação tátil da chávena? Por outras palavras, como é que os nossos inputs sensoriais se fundem numa perceção singular? Os cientistas assumiram, durante muito tempo, que os vários inputs que chegam ao neocórtex têm de convergir para um único local no cérebro, onde algo como uma chávena de café é percebido. Esta suposição faz parte da teoria da hierarquia de detetores de características. No entanto, as conexões no neocórtex não se apresentam assim. Em vez de convergirem para um único ponto, espalham-se em todas as

direções. Esta é uma das razões pelas quais o problema da ligação é considerado um mistério — mas nós propusemos uma resposta: as colunas votam. A sua perceção é o consenso a que as colunas chegam por via do voto.

Voltemos à analogia do mapa em papel. Recorde que tem um conjunto de mapas de diferentes cidades. Os mapas foram cortados em pequenos quadrados e misturados. Você é deixado num local desconhecido e vê uma cafetaria. Se encontrar cafetarias semelhantes em vários quadrados de mapa, não consegue saber onde está. Se houver cafetarias em quatro cidades diferentes, sabe que está numa dessas quatro, mas não sabe em qual.

Agora imaginemos mais quatro pessoas, iguais a si. Elas também têm mapas das cidades e são largadas na mesma cidade que você, mas em locais diferentes e aleatórios. Tal como você, não sabem em que cidade estão nem onde, exatamente. Tiram as vendas dos olhos e olham em volta. Uma pessoa vê uma biblioteca e, ao consultar os seus quadrados de mapa, encontra bibliotecas em seis cidades diferentes. Outra pessoa vê um jardim de rosas e encontra jardins de rosas em três cidades diferentes. As outras duas fazem o mesmo. Ninguém sabe em que cidade está, mas todos têm uma lista de cidades possíveis. Agora todos votam. Cada um tem uma aplicação no telemóvel que lista as cidades e localizações em que poderiam estar. Todos podem ver as listas dos outros. Apenas a Cidade 9 aparece nas listas de toda a gente; portanto, todos agora sabem que estão na Cidade 9. Ao comparar as listas de cidades possíveis e manter apenas as que aparecem em todas as listas, todos descobrem instantaneamente onde estão. A este processo chamamos votação.

Neste exemplo, as cinco pessoas são como cinco pontas de dedo a tocar diferentes locais de um objeto. Individualmente, não conseguem determinar que objeto estão a tocar, mas em conjunto conseguem. Se toca num objeto com apenas um dedo, tem de o mover para o reconhecer. Mas se o agarrar com toda a mão, normalmente consegue reconhecê-lo de imediato. Na maior parte dos casos, usar cinco dedos requer menos movimento do que usar apenas um. Do mesmo modo, se observar um objeto através de uma palhinha, tem de movê-la para o reconhecer. Mas se o vir com

o olho inteiro, normalmente reconhece o objeto sem ter de se mover.

Prosseguindo com a analogia, imagine que, entre as cinco pessoas largadas na cidade, uma só consegue ouvir. Os quadrados do seu mapa estão assinalados com os sons que deve ouvir em cada local. Quando ouve uma fonte, ou pássaros nas árvores, ou música vinda de uma cantina, encontra os quadrados de mapa onde esses sons podem ser ouvidos. Do mesmo modo, imaginemos que duas pessoas só conseguem sentir tato. Os seus mapas estão marcados com as sensações táteis que esperam sentir em diferentes locais. Finalmente, duas pessoas só conseguem ver. Os seus quadrados de mapa indicam o que esperam ver em cada ponto. Temos agora cinco pessoas com três tipos diferentes de sensores: visão, tato e audição. Todos percebem alguma coisa, mas não conseguem determinar onde estão — por isso votam. O mecanismo de votação funciona exatamente como descrevi antes. Basta que concordem sobre a cidade — os restantes detalhes não são importantes. A votação funciona entre diferentes modalidades sensoriais.

Repare que não é necessário saber muito sobre os outros. Não precisa de saber que sentidos eles têm, nem quantos mapas possuem. Não precisa de saber se os seus mapas têm mais ou menos quadrados que os seus, ou se os quadrados representam áreas maiores ou menores. Não precisa de saber como eles se movem. Talvez algumas pessoas consigam saltar sobre os quadrados e outras só se consigam mover na diagonal. Nenhum desses detalhes importa. O único requisito é que todos possam partilhar a sua lista de cidades possíveis. A votação entre colunas corticais resolve o problema da ligação. Permite ao cérebro unir numerosos tipos de input sensorial numa representação única daquilo que está a ser percebido.

Há ainda um outro elemento na votação. Quando agarra um objeto com a mão, acreditamos que as colunas táteis que representam os seus dedos partilham uma outra informação — a sua posição relativa umas às outras — o que facilita a identificação do que estão a tocar. Imagine os nossos cinco exploradores largados numa cidade desconhecida. É possível, até provável, que vejam cinco coisas que existem em várias cidades, como duas

cafetarias, uma biblioteca, um parque e uma fonte. A votação eliminará qualquer cidade que não tenha todas essas características, mas os exploradores ainda não saberão com certeza onde estão, pois, várias cidades podem conter esses cinco elementos. Contudo, se os cinco exploradores souberem qual é a sua posição relativa entre si, então poderão eliminar quaisquer cidades que não tenham essas cinco características dispostas naquela configuração específica. Suspeitamos que a informação sobre posição relativa também é partilhada entre algumas colunas corticais.

Como é que a Votação é Realizada no Cérebro?

Recorde que a maioria das conexões numa coluna cortical seguem um percurso ascendente e descendente entre as camadas, permanecendo na sua maioria dentro dos limites da própria coluna. Existem algumas exceções bem conhecidas a esta norma. As células em certas camadas enviam axónios a longas distâncias dentro do neocórtex. Podem, por exemplo, enviar os seus axónios de um lado do cérebro para o outro — entre as áreas que representam a mão esquerda e a mão direita. Ou podem enviar os seus axónios do V1, a região visual primária, até ao A1, a região auditiva primária. Propomos que estas células com conexões de longo alcance são as responsáveis pela votação.

Só faz sentido que determinadas células participem na votação. A maioria das células numa coluna não representa o tipo de informação sobre a qual as colunas poderiam votar. Por exemplo, o input sensorial de uma coluna é diferente do input sensorial de outras colunas, e por isso as células que recebem esses inputs não projetam para outras colunas. Mas as células que representam qual objeto está a ser percecionado podem votar e irão projetar essa informação de forma alargada.

A ideia básica de como as colunas podem votar não é complicada. Através das suas conexões de longo alcance, uma coluna transmite aquilo que pensa estar a observar. Muitas vezes, uma coluna está incerta — nesse caso, os seus neurónios enviam múltiplas possibilidades em simultâneo. Ao mesmo tempo, a coluna recebe projeções vindas de outras colunas, representando as suas

suposições. As hipóteses mais comuns suprimem as menos comuns, até que toda a rede se estabilize numa única resposta. Surpreendentemente, uma coluna não precisa de enviar o seu voto para todas as outras colunas. O mecanismo de votação funciona bem mesmo quando os axónios de longo alcance se conetam apenas a um subconjunto pequeno e escolhido aleatoriamente de outras colunas. A votação também exige uma fase de aprendizagem. Nos nossos artigos publicados, descrevemos simulações de software que mostram como a aprendizagem ocorre e como a votação acontece de forma rápida e fiável.

Estabilidade da Perceção

A votação entre colunas resolve outro mistério do cérebro: por que é que a nossa perceção do mundo parece estável quando os inputs dirigidos ao cérebro estão constantemente a mudar? Quando os nossos olhos fazem movimentos sacádicos, o input dirigido ao neocórtex altera-se a cada movimento ocular, e, por conseguinte, os neurónios ativos também devem mudar. No entanto, a nossa perceção visual permanece estável; o mundo não parece estar aos saltos sempre que os olhos se movem. Na maior parte do tempo, nem sequer temos consciência de que os olhos se estão a mover. Uma estabilidade de perceção semelhante ocorre com o tato. Imagine que tem uma chávena de café em cima da secretária e que a está a segurar com a mão. Percebe a presença da chávena. Agora, passe distraidamente os dedos sobre a sua superfície. À medida que o faz, os inputs dirigidos ao neocórtex mudam, mas a sua perceção é a de que a chávena permanece estável. Não lhe parece que a chávena esteja a mudar ou a mover-se.

Então por que é que a nossa perceção é estável, e por que é que não temos consciência da alteração dos inputs provenientes da pele e dos olhos? Reconhecer um objeto significa que as colunas votaram e chegaram a acordo sobre qual o objeto que estão a percecionar. Os neurónios de votação em cada coluna formam um padrão estável que representa o objeto e a sua posição relativa em relação a si. A atividade desses neurónios de votação não muda quando movimenta os olhos ou os dedos — desde que estejam a percecionar o mesmo objeto. Os outros neurónios em cada coluna

mudam com o movimento, mas os neurónios que representam o objeto não mudam.

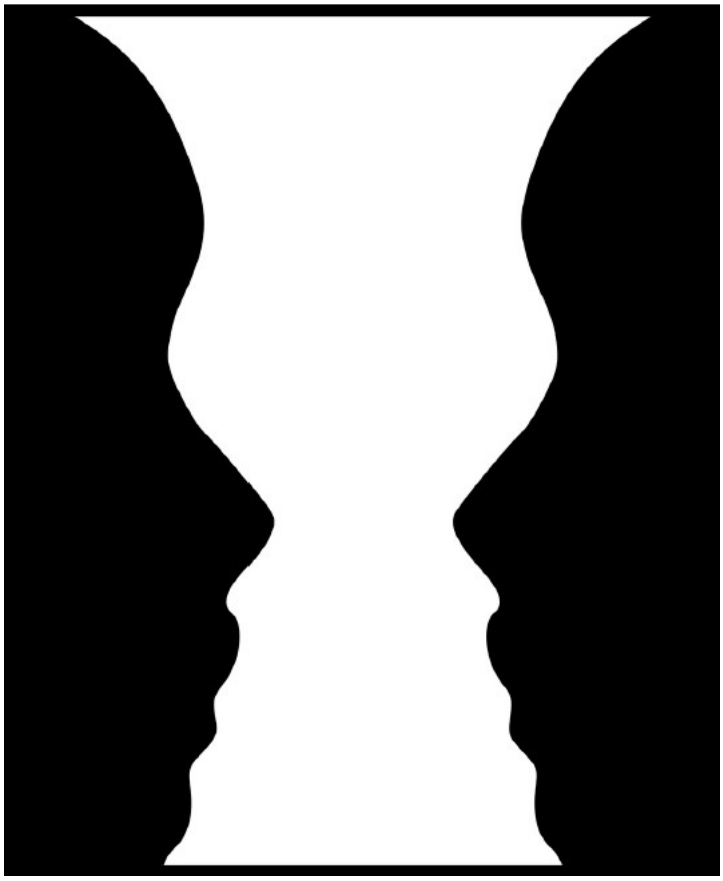
Se pudesse observar o neocórtex de cima, veria um padrão estável de atividade numa das camadas celulares. Esta estabilidade estender-se-ia por grandes áreas, abrangendo milhares de colunas. Esses são os neurónios de votação. A atividade das células noutras camadas mudaria rapidamente, de coluna para coluna. Aquilo que percecionamos baseia-se nos neurónios de votação estáveis. A informação vinda desses neurónios é disseminada amplamente para outras áreas do cérebro, onde pode ser convertida em linguagem ou armazenada na memória de curto prazo. Não temos consciência da atividade em constante mudança dentro de cada coluna, porque esta permanece confinada à própria coluna e não está acessível a outras partes do cérebro.

Para interromper crises epiléticas, os médicos por vezes cortam as conexões entre os hemisférios esquerdo e direito do neocórtex. Após essa cirurgia, os pacientes agem como se tivessem dois cérebros. As experiências demonstram claramente que os dois lados do cérebro desenvolvem pensamentos distintos e chegam a conclusões diferentes. A votação entre colunas pode explicar porquê. As conexões entre o neocórtex esquerdo e o direito são usadas para votação. Quando são cortadas, deixa de haver um mecanismo para os dois lados votarem, e por isso chegam a conclusões independentes.

O número de neurónios de votação ativos em qualquer momento é reduzido. Se fosse um cientista a observar os neurónios responsáveis pela votação, poderia ver 98% das células silenciosas e 2% em disparo contínuo. A atividade das restantes células nas colunas corticais mudaria consoante o input. Seria fácil concentrar-se nos neurónios que estão a mudar e não reparar na importância dos neurónios de votação.

O cérebro quer chegar a um consenso. Provavelmente já viu a imagem abaixo, que pode surgir como um vaso ou como dois rostos. Em exemplos como este, as colunas não conseguem determinar qual é o objeto correto. É como se tivessem dois mapas de duas cidades diferentes, mas os mapas, pelo menos em algumas zonas, fossem idênticos. A "Cidade do Vaso" e a "Cidade dos Rostos"

são semelhantes. A camada de votação quer chegar a um consenso — não permite que dois objetos estejam ativos em simultâneo —, por isso escolhe uma possibilidade em detrimento da outra. Consegue percecionar rostos ou um vaso, mas não ambos ao mesmo tempo.



Atenção

É comum os nossos sentidos estarem parcialmente obstruídos, como quando olhamos para alguém que está atrás da porta de um carro. Embora só vejamos metade da pessoa, não somos enganados. Sabemos que uma pessoa inteira está por detrás da porta. As colunas que veem a pessoa votam, e têm a certeza de que esse objeto é uma pessoa. Os neurónios de votação projetam-se nas colunas cujo input está obstruído, e agora todas as colunas sabem que há uma pessoa. Mesmo as colunas que estão bloqueadas conseguem prever o que veriam se a porta não estivesse lá.

Um momento depois, podemos deslocar a nossa atenção para a porta do carro. Tal como na imagem bistável do vaso e dos rostos, existem duas interpretações possíveis para o input. Podemos alternar a nossa atenção entre "pessoa" e "porta". A cada mudança, os neurónios de votação estabilizam num objeto diferente. Temos a perceção de que ambos os objetos estão presentes, embora só possamos atender a um de cada vez.

O cérebro pode focar-se em partes maiores ou menores de uma cena visual. Por exemplo, posso concentrar-me na porta inteira do carro ou apenas no puxador. Não se sabe exatamente como o cérebro faz isto, mas envolve uma parte do cérebro chamada tálamo, que está fortemente ligada a todas as áreas do neocórtex.

A atenção desempenha um papel essencial na forma como o cérebro aprende modelos. À medida que vai passando o dia, o seu cérebro está rápida e constantemente a focar-se em diferentes coisas. Por exemplo, ao ler, a sua atenção passa de palavra em palavra. Ou, ao olhar para um edifício, a sua atenção pode passar do edifício para uma janela, depois para a porta, para a maçaneta da porta, de volta para a porta, e assim por diante. O que pensamos que acontece é que, sempre que se foca num objeto diferente, o seu cérebro determina a localização desse objeto em relação ao objeto anterior a que prestou atenção. É automático. Faz parte do processo atencional. Por exemplo, entro numa sala de jantar. Posso primeiro prestar atenção a uma das cadeiras e depois à mesa. O meu cérebro reconhece uma cadeira e depois reconhece uma mesa. Contudo, o meu cérebro também calcula a posição relativa da cadeira em relação à mesa. À medida que olho em redor da sala de jantar, o meu cérebro não só está a reconhecer todos os objetos na sala, como simultaneamente a determinar onde cada objeto está em relação aos outros objetos e à própria sala. Com apenas um olhar em redor, o meu cérebro constrói um modelo da sala que inclui todos os objetos a que prestei atenção.

Muitas vezes, os modelos que aprende são temporários. Suponhamos que se senta para uma refeição em família na sala de jantar. Olha em redor da mesa e vê os vários pratos. Depois peço-lhe que feche os olhos e me diga onde estão as batatas. É quase certo que conseguirá fazê-lo, o que prova que aprendeu um modelo

da mesa e do seu conteúdo no curto espaço de tempo em que a observou. Alguns minutos mais tarde, depois da comida ter sido passada, peço-lhe novamente que feche os olhos e aponte para as batatas. Agora apontará para um novo local, onde as viu pela última vez. O objetivo deste exemplo é mostrar que estamos constantemente a aprender modelos de tudo o que sentimos. Se a disposição dos elementos nos nossos modelos se mantiver fixa, como o logótipo na chávena de café, então o modelo pode ser recordado por muito tempo. Se a disposição se alterar, como os pratos na mesa, então os modelos são temporários.

O neocórtex nunca deixa de aprender modelos. Cada mudança de atenção — quer esteja a olhar para os pratos na mesa de jantar, a caminhar pela rua, ou a reparar num logótipo numa chávena de café — está a adicionar mais um elemento a um modelo de algo. É o mesmo processo de aprendizagem, quer os modelos sejam efémeros ou duradouros.

Hierarquia na Teoria dos Mil Cérebros

Durante décadas, a maioria dos neurocientistas aderiu à teoria hierárquica das características, e com boas razões. Esta teoria, embora tenha muitos problemas, encaixa-se em muitos dados. A nossa teoria propõe uma forma diferente de pensar sobre o neocórtex. A Teoria dos Mil Cérebros afirma que uma hierarquia de regiões do neocórtex não é estritamente necessária. Mesmo uma única região cortical pode reconhecer objetos, como é evidenciado pelo sistema visual do rato. Então, qual das duas está correta? O neocórtex está organizado como uma hierarquia ou como milhares de modelos que votam para alcançar um consenso?

A anatomia do neocórtex sugere que ambos os tipos de conexões existem. Como podemos fazer sentido disto? A nossa teoria propõe uma forma diferente de pensar as conexões, compatível tanto com modelos hierárquicos como com modelos de coluna única. Propusemos que o que é transmitido entre níveis hierárquicos não são características, mas sim objetos completos. Em vez de o neocórtex usar a hierarquia para montar características até reconhecer um objeto, o neocórtex usa a hierarquia para montar objetos em objetos mais complexos.

Falei anteriormente da composição hierárquica. Recorda-se do exemplo de uma chávena de café com um logótipo impresso na lateral. Aprendemos um novo objeto como este ao prestar primeiro atenção à chávena e depois ao logótipo. O logótipo também é composto por objetos, como um gráfico e uma palavra, mas não precisamos de nos lembrar onde estão as características do logótipo em relação à chávena. Só precisamos de aprender a posição relativa do referencial do logótipo em relação ao referencial da chávena. Todos os detalhes do logótipo estão implicitamente incluídos.

É assim que todo o mundo é aprendido: como uma hierarquia complexa de objetos localizados relativamente a outros objetos. A forma exata como o neocórtex faz isto ainda não é clara. Por exemplo, suspeitamos que algum grau de aprendizagem hierárquica ocorre dentro de cada coluna, mas certamente não tudo. Parte disso será gerido pelas conexões hierárquicas entre regiões. Quanto está a ser aprendido dentro de uma única coluna e quanto está a ser aprendido nas conexões entre regiões é algo que ainda não compreendemos. Estamos a trabalhar neste problema. A resposta exigirá quase de certeza uma melhor compreensão da atenção, razão pela qual estamos a estudar o tálamo.

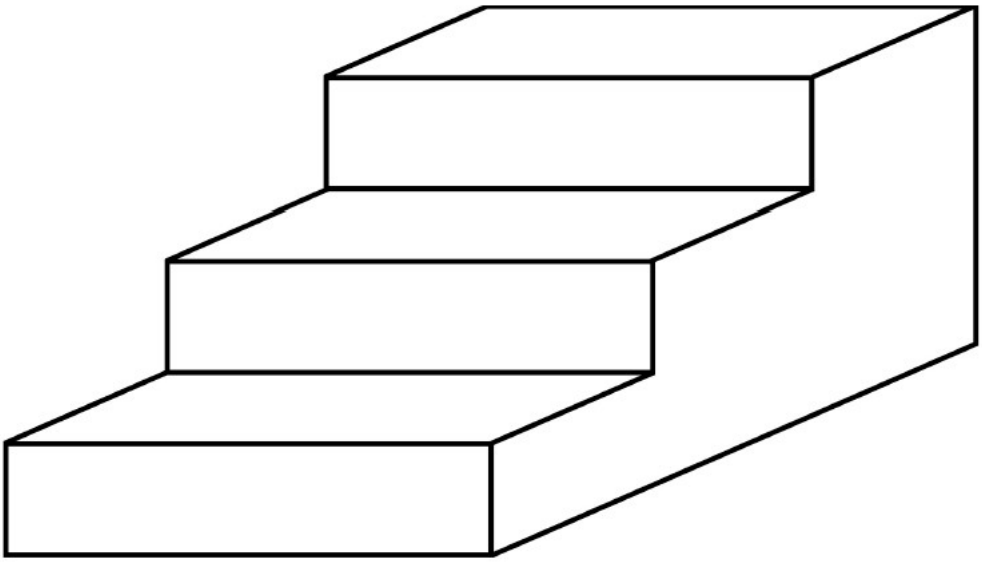
Mais cedo neste capítulo, fiz uma lista de problemas com a visão comum de que o neocórtex é uma hierarquia de detetores de características. Vamos rever essa lista, desta vez discutindo como a Teoria dos Mil Cérebros responde a cada problema, começando pelo papel essencial do movimento:

- A Teoria dos Mil Cérebros é inerentemente uma teoria sensório-motora. Explica como aprendemos e reconhecemos objetos através do movimento. E, de forma importante, também explica por que conseguimos, por vezes, reconhecer objetos sem nos movermos, como quando vemos brevemente uma imagem num ecrã ou agarramos um objeto com todos os dedos. Assim, a Teoria dos Mil Cérebros é um superconjunto do modelo hierárquico.
- O tamanho relativamente grande das regiões V1 e V2 nos primatas, e o tamanho singularmente grande da região V1 nos ratos, faz sentido na Teoria dos Mil

Cérebros porque cada coluna pode reconhecer objetos completos. Ao contrário do que muitos neurocientistas acreditam atualmente, a Teoria dos Mil Cérebros afirma que a maior parte daquilo a que chamamos visão ocorre nas regiões V1 e V2. As regiões primária e secundária relacionadas com o tato também são relativamente grandes.

- A Teoria dos Mil Cérebros consegue explicar o mistério de como certos neurónios sabem qual será o seu próximo input enquanto os olhos ainda estão em movimento. Na teoria, cada coluna tem modelos de objetos completos e, por isso, sabe o que deveria ser sentido em cada localização de um objeto. Se uma coluna conhece a localização atual do seu input e como os olhos estão a mover-se, então pode prever a nova localização e o que irá sentir aí. É o mesmo que olhar para o mapa de uma cidade e prever o que verá se começar a caminhar numa dada direção.
- O problema da ligação baseia-se na suposição de que o neocórtex tem um único modelo para cada objeto do mundo. A Teoria dos Mil Cérebros inverte esta ideia e diz que existem milhares de modelos para cada objeto. Os diversos inputs para o cérebro não estão fundidos ou combinados num único modelo. Não importa que as colunas tenham tipos diferentes de input, ou que uma coluna represente uma pequena parte da retina e a seguinte uma parte maior. Não importa se a retina tem buracos, da mesma forma que não importa que haja espaços entre os seus dedos. O padrão projetado para a região V1 pode estar distorcido e desorganizado e isso não terá importância, porque nenhuma parte do neocórtex tenta reconstituir essa representação baralhada. O mecanismo de votação da Teoria dos Mil Cérebros explica por que temos uma perceção singular e não distorcida. Também explica como o reconhecimento de um objeto num dado sentido leva a previsões noutros sentidos.

- Finalmente, a Teoria dos Mil Cérebros mostra como o neocórtex aprende modelos tridimensionais de objetos usando referenciais. Como mais uma pequena evidência, observe a imagem seguinte. É um conjunto de linhas retas impressas numa superfície plana. Não há ponto de fuga, nem linhas convergentes, nem contrastes em diminuição que sugiram profundidade. Contudo, não consegue olhar para esta imagem sem a ver como um conjunto de degraus tridimensionais. Não importa que a imagem que está a observar seja bidimensional; os modelos no seu neocórtex são tridimensionais — e é isso que percebe.



O cérebro é complexo. Os detalhes de como as células de lugar e as células de grelha criam referenciais, aprendem modelos de ambientes e planeiam comportamentos são mais complexos do que descrevi e apenas parcialmente compreendidos. Estamos a propor que o neocórtex utiliza mecanismos semelhantes, que são igualmente complexos e ainda menos compreendidos. Esta é uma área de investigação ativa, tanto para neurocientistas experimentais como para teóricos como nós.

Para avançar mais nestes e noutros tópicos, teria de introduzir detalhes adicionais de neuroanatomia e neurofisiologia — detalhes que são difíceis de descrever e não essenciais para compreender os fundamentos da Teoria dos Mil Cérebros da Inteligência. Por

conseguinte, chegámos a uma fronteira — uma fronteira onde termina aquilo que este livro explora e começa aquilo que os artigos científicos precisam de abordar.

Na introdução deste livro, disse que o cérebro é como um puzzle. Temos dezenas de milhares de factos sobre o cérebro, cada um como uma peça do puzzle. Mas, sem um quadro teórico, não fazíamos ideia de como seria a solução do puzzle. Sem um enquadramento teórico, o melhor que conseguíamos fazer era juntar algumas peças aqui e ali. A Teoria dos Mil Cérebros é um enquadramento teórico; é como completar a borda do puzzle e saber qual é a imagem geral. Enquanto escrevo, já preenchemos algumas partes do interior do puzzle, enquanto muitas outras ainda estão por fazer. Embora ainda reste muito por fazer, a nossa tarefa agora é mais simples, porque conhecer o enquadramento adequado torna mais claro que partes ainda faltam preencher.

Não quero deixar a impressão errada de que compreendemos tudo o que o neocórtex faz. Estamos longe disso. O número de coisas que não compreendemos sobre o cérebro em geral, e sobre o neocórtex em particular, é enorme. No entanto, não acredito que venha a surgir outro enquadramento teórico global, uma forma diferente de organizar as peças da borda do puzzle. Os quadros teóricos são modificados e refinados ao longo do tempo, e espero que o mesmo aconteça com a Teoria dos Mil Cérebros, mas creio que as ideias centrais que aqui apresentei permanecerão, em grande medida, intactas.

Antes de terminarmos este capítulo e a Parte 1 do livro, quero contar-lhe o resto da história sobre a vez em que conheci Vernon Mountcastle. Recorde-se que fiz uma palestra na Universidade Johns Hopkins e, no final do dia, reuni-me com Mountcastle e com o diretor do seu departamento. Chegara o momento de me ir embora; tinha um voo para apanhar. Despedimo-nos, e um carro aguardava-me à porta. Quando saía pelo gabinete, Mountcastle intercetou-me, pousou a mão no meu

ombro e disse, num tom de quem está a dar um conselho: "Devias parar de falar sobre hierarquia. Isso, na verdade, não existe."

Fiquei estupefacto. Mountcastle era o maior especialista mundial no neocórtex, e estava a dizer-me que uma das suas características mais evidentes e bem documentadas não existia. Fiquei tão surpreendido como se o próprio Francis Crick me dissesse: "Ah, essa molécula de ADN, na verdade não codifica os teus genes." Não soube o que responder, por isso não disse nada. Enquanto seguia no carro a caminho do aeroporto, tentei dar sentido às suas palavras de despedida.

Hoje, a minha compreensão da hierarquia no neocórtex mudou radicalmente — é muito menos hierárquica do que outrora pensei. Teria Vernon Mountcastle já essa noção naquela altura? Teria ele uma base teórica para afirmar que a hierarquia não existia realmente? Estaria a pensar em resultados experimentais que eu desconhecia? Morreu em 2015, e nunca mais terei oportunidade de lhe perguntar. Após a sua morte, tomei a iniciativa de reler muitos dos seus livros e artigos. O seu pensamento e escrita são sempre perspicazes. O seu *Perceptual Neuroscience: The Cerebral Cortex*, de 1998, é um livro fisicamente belo e continua a ser um dos meus favoritos sobre o cérebro. Quando penso naquele dia, percebo que teria sido sensato arriscar perder o voo pela oportunidade de conversar mais um pouco com ele. Ainda mais, gostava de poder falar com ele agora. Gosto de acreditar que teria apreciado a teoria que acabei de lhe descrever.

Agora, quero dirigir a nossa atenção para o impacto que a Teoria dos Mil Cérebros terá no nosso futuro.

PARTE 2

INTELIGÊNCIA DE MÁQUINA

No seu célebre livro *A Estrutura das Revoluções Científicas*, o historiador Thomas Kuhn argumentou que a maior parte do progresso científico assenta em quadros teóricos amplamente aceites, a que chamou paradigmas científicos. De tempos a tempos, um paradigma estabelecido é derrubado e substituído por um novo paradigma — aquilo a que Kuhn chamou uma revolução científica.

Hoje em dia, existem paradigmas estabelecidos para muitos subcampos da neurociência, como a forma como o cérebro evoluiu, as doenças relacionadas com o cérebro, e as células de grelha e de localização. Os cientistas que trabalham nestas áreas partilham terminologia e técnicas experimentais, e concordam sobre quais as questões que querem responder. Mas não existe um paradigma geralmente aceite para o neocórtex e para a inteligência. Há pouco consenso sobre o que faz o neocórtex ou sequer sobre quais as perguntas que devemos tentar responder. Kuhn diria que o estudo da inteligência e do neocórtex está num estado pré-paradigmático.

Na Parte 1 deste livro, apresentei uma nova teoria sobre o funcionamento do neocórtex e sobre o que significa ser inteligente. Pode-se dizer que estou a propor um paradigma para o estudo do neocórtex. Estou confiante de que esta teoria está em grande parte correta, mas, mais importante ainda, é uma teoria testável. Experiências em curso e futuras dirão quais as partes da teoria que estão certas e quais precisam de ser modificadas.

Nesta segunda parte do livro, vou descrever como a nossa nova teoria irá impactar o futuro da inteligência artificial. A investigação em IA tem um paradigma estabelecido: um conjunto comum de técnicas conhecidas por redes neuronais artificiais. Os cientistas da IA partilham uma terminologia e objetivos, o que tem permitido à área progredir de forma constante nos últimos anos.

A Teoria dos Mil Cérebros da Inteligência sugere que o futuro da inteligência artificial será substancialmente diferente do que a maioria dos investigadores de IA atualmente imagina. Acredito que a IA está pronta para uma revolução científica, e que os princípios da inteligência que descrevi anteriormente serão a base dessa revolução.

Escrevo isto com alguma hesitação, por causa de uma experiência que tive no início da minha carreira, quando falei sobre o futuro da computação. As coisas não correram bem.

Pouco depois de ter fundado a Palm Computing, fui convidado a dar uma palestra na Intel. Uma vez por ano, a Intel reunia várias centenas dos seus quadros superiores no Vale do Silício, durante três dias, para uma reunião de planeamento. Como parte dessas reuniões, convidavam algumas pessoas de fora para falar a todo o grupo, e em 1992 fui um desses oradores. Considerei isso uma honra. A Intel liderava a revolução dos computadores pessoais e era uma das empresas mais respeitadas e poderosas do mundo. A minha empresa, a Palm, era uma pequena start-up que ainda não tinha lançado o seu primeiro produto. A minha palestra era sobre o futuro da computação pessoal.

Propus que o futuro da computação pessoal iria ser dominado por computadores suficientemente pequenos para caberem no bolso. Esses dispositivos custariam entre quinhentos e mil dólares e funcionariam durante todo o dia com uma só carga de bateria. Para milhares de milhões de pessoas em todo o mundo, um computador de bolso seria o único que alguma vez possuiriam. Para mim, esta transição era inevitável. Milhares de milhões de pessoas queriam acesso a computadores, mas os portáteis e os computadores de secretária eram demasiado caros e difíceis de usar. Via uma força inexorável a puxar no sentido dos computadores de bolso, mais fáceis de usar e menos dispendiosos.

Na altura, havia centenas de milhões de computadores pessoais, entre portáteis e de secretária. A Intel vendia os processadores para a maioria deles. Cada processador custava cerca de quatrocentos dólares e consumia demasiada energia para poder ser usado num computador de bolso alimentado a bateria. Sugeri aos gestores da Intel que, se quisessem manter a sua

posição de liderança na computação pessoal, deviam concentrar-se em três áreas: reduzir o consumo de energia, tornar os seus chips mais pequenos, e descobrir como obter lucro com um produto que se vendesse por menos de mil dólares. O tom da minha palestra era modesto, nada dogmático. Era como quem diz: "Ah, já agora, creio que isto vai acontecer, e talvez seja útil considerarem estas implicações."

Quando terminei a palestra, responderia a algumas perguntas do público. Todos estavam sentados à mesa de almoço e a comida só seria servida depois da minha intervenção, por isso não esperava muitas perguntas. Lembro-me de apenas uma. Uma pessoa levantou-se e perguntou, num tom que me pareceu ligeiramente trocista: "Para que é que as pessoas vão usar esses computadores de bolso?" Foi difícil responder.

Na altura, os computadores pessoais eram usados sobretudo para processamento de texto, folhas de cálculo e bases de dados. Nenhuma destas aplicações se adequava a um computador de bolso com ecrã pequeno e sem teclado. A lógica dizia-me que os computadores de bolso seriam usados sobretudo para aceder à informação, não para a criar — e foi isso que respondi. Disse que as primeiras aplicações seriam os calendários e listas de contatos, mas sabia que isso não chegaria para transformar a computação pessoal. Acrescentei que iríamos descobrir novas aplicações que se tornariam mais importantes.

Lembre-se que, no início de 1992, não havia música digital, nem fotografia digital, nem Wi-Fi, nem Bluetooth, nem dados nos telemóveis. O primeiro navegador da web para consumidores ainda não tinha sido inventado. Não fazia ideia de que essas tecnologias viriam a ser criadas, por isso não podia imaginar aplicações baseadas nelas. Mas sabia que as pessoas queriam sempre mais informação, e que, de algum modo, descobriríamos como entregá-la a computadores móveis.

Depois da palestra, sentei-me a uma mesa com o Dr. Gordon Moore, o lendário fundador da Intel. Era uma mesa redonda, com cerca de dez pessoas. Perguntei ao Dr. Moore o que tinha achado da minha intervenção. Todos se calaram para ouvir a resposta. Ele evitou dar-me uma resposta direta e, a partir daí, evitou falar

comigo durante o resto da refeição. Depressa se tornou evidente que nem ele nem mais ninguém na mesa acreditavam no que eu dissera.

Fiquei abalado com esta experiência. Se nem sequer conseguia que as pessoas mais inteligentes e bem-sucedidas da informática considerassem a minha proposta, talvez estivesse errado, ou talvez a transição para os computadores de bolso fosse muito mais difícil do que eu imaginava. Decidi então que o melhor caminho para mim seria concentrar-me em construir computadores de bolso, em vez de me preocupar com o que os outros acreditavam. A partir desse dia, evitei dar palestras "visionárias" sobre o futuro da informática e dediquei-me, tanto quanto possível, a tornar esse futuro realidade.

Hoje, dou por mim numa situação semelhante. Daqui para a frente, vou descrever um futuro que é diferente do que a maioria das pessoas — e, na verdade, a maioria dos especialistas — está à espera. Primeiro, descrevo um futuro da inteligência artificial que contraria o pensamento atual da maioria dos líderes da área, e depois, na Parte 3, apresento uma visão do futuro da humanidade que provavelmente nunca considerou. É claro que posso estar enganado; prever o futuro é notoriamente difícil. Mas, para mim, as ideias que estou prestes a apresentar parecem inevitáveis, mais como deduções lógicas do que como especulações. No entanto, como experienciei na Intel há muitos anos, talvez não consiga convencer toda a gente. Farei o meu melhor e peço-lhe que mantenha a mente aberta.

Nos próximos quatro capítulos, falo sobre o futuro da inteligência artificial. A IA está atualmente a passar por um renascimento. É uma das áreas mais quentes da tecnologia. Todos os dias surgem novas aplicações, novos investimentos e melhorias de desempenho. A área da IA é dominada pelas redes neuronais artificiais, embora estas não se pareçam em nada com as redes de neurónios do cérebro. Vou defender que o futuro da IA se baseará em princípios diferentes dos atuais — princípios que imitam mais de perto o funcionamento do cérebro. Para construirmos máquinas verdadeiramente inteligentes, temos de concebê-las segundo os princípios que expus na primeira parte do livro.

Não sei quais serão as futuras aplicações da IA. Mas, tal como a transição da computação pessoal para os dispositivos móveis, vejo a mudança da IA para princípios inspirados no cérebro como inevitável.

CAPÍTULO 8

Por Que É Que Não Existe "Eu" na IA

Desde a sua génese, em 1956, o campo da inteligência artificial passou por vários ciclos de entusiasmo seguidos de desilusão. Os cientistas da área designam estas fases como "verões da IA" e "invernos da IA". Cada vaga foi impulsionada por uma nova tecnologia que prometia colocar-nos no caminho da criação de máquinas inteligentes, mas que, em última análise, se revelou insuficiente. Atualmente, a IA vive mais uma vaga de entusiasmo — mais um verão da IA — e, uma vez mais, as expetativas no setor são elevadas. O conjunto de tecnologias que impulsiona este novo surto são as redes neuronais artificiais, frequentemente denominadas *deep learning (aprendizagem profunda)*. Estes métodos alcançaram resultados notáveis em tarefas como rotular imagens, reconhecer linguagem falada e conduzir automóveis. Em 2011, um computador venceu os melhores concorrentes humanos no concurso televisivo *Jeopardy*!, e em 2016, outro computador derrotou o melhor jogador do mundo no jogo *Go*. Estes dois feitos foram manchete em todo o mundo. São, sem dúvida, impressionantes — mas será que alguma destas máquinas é verdadeiramente inteligente?

A maioria das pessoas, incluindo a maior parte dos investigadores em IA, não pensa assim. Existem inúmeras formas pelas quais a inteligência artificial atual fica aquém da inteligência humana. Por exemplo, os seres humanos aprendem de forma contínua. Como descrevi anteriormente, estamos constantemente a corrigir e a ampliar o nosso modelo do mundo. Em contraste, as redes de aprendizagem profunda têm de ser integralmente treinadas antes de poderem ser utilizadas. E, uma vez implementadas, não conseguem aprender novas informações em tempo real. Por exemplo, se quisermos ensinar a uma rede neuronal de visão a reconhecer um novo objeto, ela terá de ser totalmente

retreinada desde o início — um processo que pode levar dias. No entanto, a principal razão pela qual os sistemas atuais de IA não são considerados inteligentes é o facto de conseguirem executar apenas uma tarefa, ao passo que os humanos são capazes de realizar muitas. Em outras palavras, os sistemas de IA não são flexíveis. Qualquer ser humano, como você ou eu, pode aprender a jogar Go, cultivar a terra, programar software, pilotar um avião ou tocar um instrumento musical. Ao longo da vida, aprendemos milhares de competências, e, embora não sejamos necessariamente exímios em todas elas, possuímos uma flexibilidade notável naquilo que somos capazes de aprender. Os sistemas de IA baseados em aprendizagem profunda praticamente não exibem qualquer flexibilidade. Um computador especializado em jogar Go pode ultrapassar qualquer humano nesse jogo, mas é incapaz de fazer mais nada. Um carro autónomo pode conduzir com mais segurança do que qualquer condutor humano, mas não sabe jogar Go nem mudar um pneu furado.

O objetivo a longo prazo da investigação em IA é criar máquinas que manifestem uma inteligência semelhante à humana — máquinas capazes de aprender rapidamente novas tarefas, de reconhecer analogias entre problemas distintos e de resolver novos desafios com flexibilidade. Este objetivo é designado por *"inteligência artificial geral"*, *ou AGI (artificial general intelligence)*, para distingui-lo da IA limitada atual. A grande questão que se coloca hoje à indústria da IA é esta: estaremos, de facto, no caminho para criar máquinas verdadeiramente inteligentes, dotadas de AGI, ou iremos mais uma vez estagnar e mergulhar noutro inverno da IA? A presente vaga de IA atraiu milhares de investigadores e milhares de milhões de dólares em investimentos. Praticamente todo este capital humano e financeiro está a ser canalizado para o aperfeiçoamento das tecnologias de aprendizagem profunda. Será que este investimento nos conduzirá à criação de máquinas com inteligência ao nível da humana, ou serão estas tecnologias intrinsecamente limitadas, forçando-nos, mais uma vez, a reinventar o campo da IA? Quando estamos no meio de uma bolha, é fácil deixarmo-nos contagiar pelo entusiasmo e acreditar que ele durará para sempre. No entanto, a história aconselha prudência.

Não sei durante quanto tempo continuará a crescer a presente vaga de IA. Mas sei que a aprendizagem profunda, por si só, não nos coloca no caminho certo para criar máquinas verdadeiramente inteligentes. Não chegaremos à inteligência artificial geral apenas fazendo mais do que temos feito até agora. Teremos de seguir uma via diferente.

Duas Vias para a AGI

Existem duas vias que os investigadores em Inteligência Artificial têm seguido na tentativa de criar máquinas inteligentes. Uma dessas vias — aquela que seguimos atualmente — consiste em fazer com que os computadores superem os seres humanos em tarefas específicas, como jogar Go ou detetar células cancerígenas em imagens médicas. A esperança é que, se conseguirmos que os computadores superem os humanos em algumas tarefas difíceis, eventualmente descobriremos como torná-los melhores do que os humanos em todas as tarefas. Nesta abordagem à IA, pouco importa como o sistema funciona, ou se o computador é flexível. O que importa é apenas que o sistema execute uma tarefa concreta melhor do que os outros sistemas de IA — e, por fim, melhor do que o melhor humano. Por exemplo, se o melhor computador a jogar Go estivesse classificado em sexto lugar no mundo, isso não teria feito manchetes — e talvez até fosse visto como um fracasso. Mas o facto de ter vencido o melhor jogador humano do mundo foi encarado como um avanço significativo.

A segunda via para criar máquinas inteligentes dá primazia à flexibilidade. Nesta abordagem, não é necessário que a IA supere os humanos — o objetivo é criar máquinas capazes de fazer muitas coisas e de aplicar o que aprenderam numa tarefa a outras diferentes. O êxito, neste caminho, pode ser alcançado com uma máquina que possua as capacidades de uma criança de cinco anos — ou mesmo de um cão. A esperança é que, se conseguirmos primeiro compreender como construir sistemas de IA flexíveis, então, com essa base, poderemos eventualmente desenvolver sistemas que igualem ou superem os humanos.

Esta segunda via foi preferida em algumas das vagas iniciais da IA. Contudo, revelou-se demasiado difícil. Os cientistas

perceberam que possuir as capacidades de uma criança de cinco anos exige um volume enorme de conhecimento quotidiano. As crianças sabem milhares de coisas sobre o mundo. Sabem que os líquidos se entornam, que as bolas rolam e que os cães ladram. Sabem usar lápis, marcadores, papel e cola. Sabem abrir livros, e que o papel pode rasgar-se. Conhecem milhares de palavras e sabem usá-las para influenciar o comportamento dos outros. Os investigadores em IA não conseguiram descobrir como programar este conhecimento quotidiano num computador — nem como fazer com que o computador o aprendesse por si.

A parte difícil do conhecimento não está em enunciar um facto, mas sim em representá-lo de forma útil. Tomemos, por exemplo, a frase "As bolas são redondas." Uma criança de cinco anos compreende o que isso significa. Podemos facilmente introduzir essa frase num computador — mas como poderá ele compreendê-la? As palavras "bola" e "redonda" têm vários significados. Uma "bola" pode ser um baile, que não é redondo, e uma pizza é redonda, mas não como uma bola. Para compreender "bola", o computador teria de associar a palavra a diferentes significados — e cada significado a diferentes relações com outras palavras. Os objetos também têm ações associadas. Algumas bolas saltam, mas bolas de futebol saltam de maneira diferente das de basebol, que por sua vez saltam de forma distinta das de ténis. Você e eu aprendemos estas diferenças pela simples observação. Ninguém nos tem de explicar como as bolas saltam — basta lançá-las ao chão e ver o que acontece. Nem sequer temos consciência da forma como este conhecimento é armazenado no nosso cérebro. Aprender esse tipo de conhecimento quotidiano — como uma bola salta — é algo natural e sem esforço.

Os cientistas da IA não conseguiram descobrir como replicar este processo num computador. Criaram estruturas de software chamadas *schemas* e *frames* para organizar o conhecimento, mas qualquer que fosse a abordagem, acabavam sempre com um sistema caótico e inutilizável. O mundo é complexo; a quantidade de coisas que uma criança sabe — e o número de ligações entre essas coisas — parece vastamente desproporcionado. Pode soar simples, mas ninguém conseguiu descobrir como fazer um

computador saber algo tão aparentemente elementar como o que é uma bola.

Este problema é conhecido como o problema da representação do conhecimento. Alguns cientistas de IA chegaram à conclusão de que este não era apenas um grande problema da área — mas o problema. Alegaram que não poderíamos criar máquinas verdadeiramente inteligentes sem antes resolver a questão de como representar o conhecimento do dia-a-dia num sistema computacional.

As redes neuronais de aprendizagem profunda atuais não possuem conhecimento. Um computador que joga Go não sabe que Go é um jogo. Não conhece a história do jogo. Não sabe se está a jogar contra um computador ou contra um humano, nem o que significam os conceitos de "computador" ou "humano". Da mesma forma, uma rede de aprendizagem profunda que rotula imagens pode identificar uma imagem como sendo um gato — mas o computador tem um conhecimento extremamente limitado sobre gatos. Não sabe que os gatos são animais, que têm caudas, patas ou pulmões. Não sabe nada sobre pessoas que preferem gatos em vez de cães, nem que os gatos ronronam ou largam pêlo. Tudo o que a rede faz é detetar que uma imagem nova se assemelha a outras que já tinham sido rotuladas como "gato". Não há ali verdadeiro conhecimento sobre gatos.

Mais recentemente, os cientistas da IA tentaram uma abordagem diferente para codificar conhecimento. Criam redes neuronais artificiais de grande escala e treinam-nas com enormes quantidades de texto: todas as palavras de dezenas de milhares de livros, a totalidade da Wikipédia, e praticamente todo o conteúdo textual da internet. O texto é introduzido na rede palavra a palavra. Com este tipo de treino, as redes aprendem a calcular a probabilidade de certas palavras surgirem após outras. Estas redes de linguagem conseguem fazer coisas surpreendentes. Por exemplo, se lhes dermos algumas palavras, são capazes de escrever um parágrafo relacionado com essas palavras. Por vezes, é difícil perceber se o texto foi escrito por um humano ou pela rede neuronal.

Os cientistas de IA divergem quanto à questão de saber se estas redes de linguagem possuem verdadeiro conhecimento, ou se estão apenas a imitar os humanos, memorizando as estatísticas de milhões de palavras. Eu não acredito que qualquer rede de deep learning consiga alcançar o objetivo da AGI se não for capaz de modelar o mundo da forma como o cérebro o faz. As redes neuronais de aprendizagem profunda funcionam bem — mas não porque tenham resolvido o problema da representação do conhecimento. Funcionam bem precisamente porque o evitaram, confiando apenas em estatísticas e quantidades massivas de dados. O modo como estas redes funcionam é engenhoso, o seu desempenho é impressionante e o seu valor comercial é elevado. O que quero sublinhar é apenas que não possuem conhecimento — e, por isso, não estão no caminho para atingir as capacidades de uma criança de cinco anos.

O Cérebro como Modelo para a IA

Desde o momento em que comecei a interessar-me pelo estudo do cérebro, senti que seria necessário compreendê-lo a fundo antes de podermos criar máquinas inteligentes. Isto pareceu-me evidente, já que o cérebro é a única coisa que conhecemos que manifesta inteligência. Ao longo das décadas seguintes, nada me fez mudar de opinião. Essa é uma das razões pelas quais me dediquei persistentemente à teoria do cérebro: considero que é um passo inicial indispensável para a criação de uma Inteligência Artificial verdadeiramente inteligente. Vivi várias vagas de entusiasmo em torno da IA, e em cada uma delas resisti à tentação de me juntar à corrente. Para mim, era claro que as tecnologias então usadas não se pareciam minimamente com o cérebro — e, por isso, a IA acabaria por estagnar. Descobrir como o cérebro funciona é difícil, sim, mas é um passo necessário para se criar máquinas inteligentes.

Na primeira parte deste livro, descrevi os avanços que fizemos na compreensão do cérebro. Expliquei como o neocórtex aprende modelos do mundo através de referenciais espaciais semelhantes a mapas. Tal como um mapa em papel representa o conhecimento sobre uma área geográfica — como uma cidade ou país —, os mapas

no cérebro representam conhecimento sobre os objetos com que interagimos (como bicicletas ou smartphones), conhecimento sobre o nosso corpo (como a localização e o movimento dos nossos membros) e conhecimento sobre conceitos abstratos (como a matemática).

A Teoria dos Mil Cérebros resolve o problema da representação do conhecimento. Aqui fica uma analogia para ajudar a compreender como. Imaginemos que quero representar o conhecimento sobre um objeto comum, um agrafador. Os primeiros investigadores em IA tentariam fazer isto listando os nomes das diferentes partes do agrafador e descrevendo o que cada parte faz. Poderiam escrever uma regra sobre agrafadores que dissesse: "Quando a parte superior do agrafador é pressionada, sai um agrafo de uma das extremidades." Mas, para compreender esta afirmação, seria necessário definir termos como "superior", "extremidade" e "agrafo", bem como ações como "pressionada" e "sai". Além disso, essa regra, por si só, seria insuficiente: não indicaria em que direção sai o agrafo, o que acontece a seguir ou o que fazer se o agrafo ficar preso. Assim, os investigadores teriam de escrever regras adicionais. Este método de representação do conhecimento conduzia a uma lista interminável de definições e regras. Os investigadores de IA não conseguiam encontrar uma forma funcional de o pôr em prática. Os críticos argumentavam que, mesmo que todas as regras pudessem ser definidas, o computador, ainda assim, não "saberia" o que é um agrafador.

O cérebro adota uma abordagem completamente diferente para armazenar conhecimento sobre um agrafador: ele aprende um modelo. O modelo é a encarnação do conhecimento. Imagine, por um momento, que tem um pequeno agrafador dentro da cabeça. É exatamente igual a um agrafador real — tem a mesma forma, as mesmas partes, e move-se da mesma maneira — apenas em tamanho reduzido. Este modelo em miniatura representa tudo o que sabe sobre agrafadores, sem ser necessário atribuir um nome a qualquer das partes. Se quiser recordar o que acontece quando se pressiona a parte superior do agrafador, basta "pressionar" o modelo e observar o que ocorre.

Claro que não existe um agrafador físico dentro da sua cabeça. Mas as células do neocórtex aprendem um modelo virtual, que cumpre a mesma função. À medida que interage com um agrafador real, o cérebro constrói o seu modelo virtual, que inclui tudo o que observou acerca do objeto — desde a sua forma até ao seu comportamento quando o utiliza. O seu conhecimento sobre agrafadores está incorporado nesse modelo. Não existe no seu cérebro uma lista de factos ou regras sobre agrafadores.

Imaginemos agora que lhe pergunto o que acontece quando se pressiona a parte superior de um agrafador. Para responder, não vai procurar a regra correspondente e recitá-la. O seu cérebro imagina o ato de pressionar o agrafador, e o modelo recorda o que acontece. Pode usar palavras para mo descrever, mas o conhecimento não está armazenado em palavras ou regras. O conhecimento é o modelo.

Acredito que o futuro da IA será baseado nos princípios do cérebro. As verdadeiras máquinas inteligentes — AGI — aprenderão modelos do mundo através de referenciais espaciais, tal como faz o neocórtex. Considero isto inevitável. Não creio que exista outro caminho para criar máquinas verdadeiramente inteligentes.

Da Inteligência Artificial Dedicada à Inteligência Artificial Universal

A situação em que nos encontramos hoje faz-me recordar os primórdios da informática. A palavra "computador" referia-se originalmente a pessoas cujo trabalho consistia em efetuar cálculos matemáticos. Para criar tabelas numéricas ou decifrar mensagens encriptadas, dezenas de computadores humanos realizavam os cálculos necessários à mão. Os primeiros computadores eletrónicos foram concebidos para substituir os computadores humanos numa tarefa específica. Por exemplo, a melhor solução automatizada para descodificação de mensagens era uma máquina exclusivamente dedicada a esse fim. Pioneiros da computação, como Alan Turing, defenderam a criação de computadores "universais": máquinas eletrónicas que pudessem ser programadas para executar qualquer

tarefa. No entanto, na altura, ninguém sabia ainda qual seria a melhor forma de construir um computador assim.

Houve, então, um período de transição, durante o qual os computadores foram construídos em diversas formas. Havia máquinas desenhadas para tarefas específicas, computadores analógicos e computadores que só podiam ser reutilizados mediante reconfiguração manual da cablagem. Havia ainda computadores que operavam com números decimais em vez de binários. Hoje, praticamente todos os computadores seguem o modelo universal previsto por Turing. Chamamos-lhes, por isso, "máquinas de Turing universais". Com o software certo, os computadores atuais podem ser aplicados a quase qualquer tarefa. Foi o mercado que determinou que os computadores universais, de uso generalista, eram o caminho a seguir. Isto apesar de, ainda hoje, determinadas tarefas poderem ser executadas mais rapidamente ou com menor consumo energético através de soluções especializadas, como chips personalizados. Contudo, os engenheiros e projetistas de produto tendem a preferir o custo mais baixo e a conveniência dos computadores de uso geral, mesmo sabendo que uma máquina dedicada poderia ser mais rápida e eficiente.

Algo semelhante ocorrerá com a inteligência artificial. Atualmente, estamos a construir sistemas de IA especializados, concebidos para executar, da melhor forma possível, a tarefa para que foram projetados. Mas no futuro, a maioria das máquinas inteligentes será universal: mais semelhantes aos humanos, capazes de aprender praticamente qualquer coisa.

Hoje em dia, os computadores existem em muitas formas e tamanhos — desde o microcomputador de uma torradeira até aos computadores gigantes usados para simular o clima. Apesar das diferenças em tamanho e velocidade, todos estes computadores operam segundo os mesmos princípios estabelecidos por Turing e outros há muitas décadas. Todos são variantes da máquina de Turing universal.

Do mesmo modo, as máquinas inteligentes do futuro apresentar-se-ão em diversos formatos e escalas, mas quase todas funcionarão com base num conjunto comum de princípios. A maioria será composta por máquinas de aprendizagem universais,

semelhantes ao cérebro. (Matemáticos demonstraram que existem certos problemas que são irresolúveis, mesmo em teoria. Portanto, para sermos rigorosos, não existem soluções verdadeiramente "universais". Mas esta é uma questão altamente teórica, que não precisamos de considerar para os propósitos deste livro.)

Alguns investigadores argumentam que as redes neuronais atuais já são universais. Uma rede neuronal pode ser treinada para jogar Go ou para conduzir um carro. Contudo, a mesma rede neuronal não pode fazer ambas as coisas. Além disso, é necessário ajustá-la e modificá-la para que execute qualquer tarefa nova. Quando uso os termos "universal" ou "de uso generalista", refiro-me a algo mais semelhante a nós próprios: uma máquina capaz de aprender diversas coisas sem ter de apagar a memória e recomeçar do zero.

Há duas razões principais pelas quais a IA fará a transição das soluções especializadas atuais para soluções mais universais, que dominarão o futuro. A primeira é semelhante à razão pela qual os computadores universais suplantaram os dedicados: no final, são mais eficazes em termos de custo, o que impulsionou uma evolução mais rápida da tecnologia. À medida que mais pessoas usam os mesmos projetos, mais esforços são canalizados para aperfeiçoar os modelos mais populares e os ecossistemas que os suportam, conduzindo a rápidas melhorias de desempenho e redução de custos. Esta foi a força motriz por detrás do crescimento exponencial da capacidade de computação que moldou a indústria e a sociedade na segunda metade do século XX. A segunda razão é que algumas das aplicações mais importantes da inteligência artificial no futuro exigirão a flexibilidade das soluções universais. Estas aplicações terão de lidar com problemas imprevistos e criar soluções inovadoras — algo que os atuais sistemas especializados de aprendizagem profunda não conseguem fazer.

Imaginemos dois tipos de robôs. O primeiro pinta automóveis numa fábrica. Pretendemos que esses robôs sejam rápidos, precisos e consistentes. Não queremos que inventem novas técnicas de pintura todos os dias ou que questionem por que estão a pintar carros. Nestes casos, robôs dedicados e sem inteligência geral são ideais. Agora imaginemos que queremos enviar uma

equipa de robôs construtores para Marte, a fim de edificarem um habitat habitável para humanos. Esses robôs terão de utilizar diversas ferramentas e montar estruturas num ambiente não estruturado. Deparar-se-ão com obstáculos imprevistos e terão de improvisar em conjunto, adaptando soluções e ajustando os planos. Os humanos conseguem lidar com este tipo de situações, mas nenhuma máquina atual está sequer próxima de o conseguir. Estes robôs para construção em Marte terão de possuir inteligência generalista.

Poderá pensar que a necessidade de máquinas inteligentes de uso geral será limitada, e que a maioria das aplicações da IA continuará a ser abordada com tecnologias especializadas, como acontece atualmente. Mas também se pensava o mesmo em relação aos computadores de uso geral. Argumentava-se que a sua utilidade comercial estaria restrita a poucas aplicações de elevado valor. E, no entanto, sucedeu o contrário. Graças à enorme redução de custos e de tamanho, os computadores generalistas tornaram-se uma das tecnologias mais importantes e economicamente impactantes do século passado. Acredito que o mesmo acontecerá com a inteligência artificial: a IA de uso geral acabará por dominar o panorama da inteligência artificial na segunda metade do século XXI. Na década de 1940 e início dos anos 50, quando os computadores comerciais começaram a surgir, era impossível imaginar para que seriam utilizados em 1990 ou 2000. Hoje, a nossa imaginação está, de novo, limitada da mesma forma. Ninguém pode prever como serão usadas as máquinas inteligentes dentro de cinquenta ou sessenta anos.

Quando é que Algo é Inteligente?

Quando devemos considerar uma máquina inteligente? Haverá um conjunto de critérios que possamos utilizar? Esta questão é análoga a perguntar: quando é que uma máquina é um computador de uso geral? Para que uma máquina possa ser classificada como um computador de uso geral — isto é, uma máquina de Turing universal — ela tem de possuir certos componentes, como memória, uma unidade central de processamento (CPU) e software. Não é possível detetar esses elementos apenas observando o

exterior da máquina. Por exemplo, não consigo saber se o meu forno de torradas contém um computador de uso geral no seu interior, ou se tem apenas um chip personalizado. Quanto mais funcionalidades o forno possuir, mais provável será conter um computador generalista — mas a única forma segura de saber é olhar para dentro e ver como funciona.

Do mesmo modo, para que algo possa ser classificado como inteligente, é necessário que funcione com base num conjunto de princípios estruturais. Não é possível detetar se um sistema utiliza esses princípios apenas pela sua observação externa. Por exemplo, se vejo um carro a circular numa autoestrada, não consigo dizer se está a ser conduzido por um ser humano inteligente, que aprende e se adapta enquanto conduz, ou por um simples controlador que apenas mantém o carro entre duas linhas. Quanto mais complexo for o comportamento do carro, maior a probabilidade de estar a ser controlado por um agente inteligente. Mas a única forma segura de saber é examinar o seu funcionamento interno.

Então, existirá um conjunto de critérios que uma máquina deve reunir para ser considerada inteligente? Acredito que sim. A proposta que apresento sobre o que qualifica uma máquina como inteligente baseia-se no cérebro. Cada um dos quatro atributos que se seguem corresponde a algo que sabemos que o cérebro faz — e que, na minha opinião, uma máquina verdadeiramente inteligente também deverá ser capaz de fazer. Descreverei cada atributo, explicando a sua natureza, a sua importância e a forma como o cérebro o implementa. Naturalmente, as máquinas inteligentes não terão de implementar estes atributos exatamente da mesma forma que o cérebro. Por exemplo, não precisam de ser compostas por células vivas.

Nem todos concordarão com a minha escolha de atributos. Pode argumentar-se, com razão, que omiti algumas capacidades importantes. E está tudo bem. Vejo esta lista como um mínimo essencial, uma linha de base para aquilo que podemos considerar como inteligência geral artificial (AGI). Pouquíssimos sistemas de IA atuais possuem sequer um destes atributos.

1. Aprender Continuamente

O que é? A cada momento em que estamos despertos, estamos a aprender. A duração com que retemos aquilo que aprendemos varia. Algumas coisas são esquecidas rapidamente, como a disposição dos pratos sobre a mesa ou a roupa que usámos ontem. Outras permanecem connosco para o resto da vida. Aprender não é um processo separado do sentir e do agir — é um processo contínuo.

Por que é importante? O mundo está em constante mutação; por isso, o nosso modelo do mundo tem de aprender continuamente para refletir essa mudança. A maioria dos sistemas de IA atuais não aprende continuamente. Passam por um longo processo de treino e, uma vez concluído, são implementados. Esta é uma das razões pelas quais não são flexíveis. A flexibilidade exige uma adaptação contínua a condições mutáveis e a novos conhecimentos.

Como o faz o cérebro? O elemento mais importante no modo como o cérebro aprende de forma contínua é o neurónio. Quando um neurónio aprende um novo padrão, forma novas sinapses num dos ramos dendríticos. Estas novas sinapses não interferem com sinapses já existentes noutros ramos, resultantes de aprendizagens anteriores. Assim, aprender algo novo não obriga o neurónio a esquecer ou modificar aquilo que já aprendeu. Os neurónios artificiais usados nos sistemas de IA atuais não possuem esta capacidade, o que explica, em parte, por que não conseguem aprender continuamente.

2. Aprender através do Movimento

O que é? Aprendemos através do movimento. À medida que nos movemos ao longo do dia — movimentando o corpo, os membros, os olhos — esses movimentos são parte integrante do processo de aprendizagem.

Por que é importante? A inteligência requer a aprendizagem de um modelo do mundo. Não podemos apreender tudo de uma vez só; por isso, o movimento é essencial para a aprendizagem. Não se pode aprender o

modelo de uma casa sem se mover de divisão em divisão, tal como não se pode aprender a usar uma nova aplicação num smartphone sem interagir com ela. Estes movimentos não têm de ser físicos: o princípio do aprender através do movimento também se aplica a conceitos abstratos, como a matemática, ou a espaços virtuais, como a internet.

Como o faz o cérebro? A unidade de processamento no neocórtex é a coluna cortical. Cada coluna é um sistema sensório-motor completo — recebe estímulos e pode gerar comportamentos. A cada movimento, a coluna prevê qual será o seu próximo estímulo. Prever é o modo como a coluna testa e atualiza o seu modelo.

3. Muitos Modelos

O que é? O neocórtex é composto por dezenas de milhares de colunas corticais, e cada coluna aprende modelos de objetos. O conhecimento sobre qualquer coisa — por exemplo, uma chávena de café — encontra-se distribuído por múltiplos modelos complementares.

Por que é importante? Este projeto de múltiplos modelos confere flexibilidade. Ao adotarem esta arquitetura, os criadores de IA podem integrar facilmente vários tipos de sensores — como visão e tato, ou até sensores novos, como radar — e podem criar máquinas com corpos diversos. Tal como o neocórtex, o "cérebro" de uma máquina inteligente consistirá em múltiplos elementos quase idênticos que podem ser conetados a diferentes sensores móveis.

Como o faz o cérebro? A chave para que este projeto funcione é o voto entre colunas. Cada coluna opera de forma relativamente independente, mas as conexões de longo alcance no neocórtex permitem que as colunas votem sobre qual é o objeto que estão a percecionar.

4. Usar Quadros de Referência para Armazenar Conhecimento

O que é? No cérebro, o conhecimento é armazenado em quadros de referência. Estes quadros são igualmente

utilizados para fazer previsões, criar planos e realizar movimentos. Pensar consiste em ativar, sucessivamente, localizações num quadro de referência, e a informação associada a cada localização é assim recuperada.

Por que é importante? Para ser inteligente, uma máquina precisa de aprender um modelo do mundo. Esse modelo tem de incluir as formas dos objetos, como estes mudam com a interação e onde estão localizados em relação uns aos outros. Os quadros de referência são necessários para representar este tipo de informação — são a espinha dorsal do conhecimento.

Como o faz o cérebro? Cada coluna cortical estabelece o seu próprio conjunto de quadros de referência. Os autores propõem que as colunas corticais criam esses quadros através de células equivalentes às células de grelha e células de lugar, descobertas no hipocampo e associadas à navegação espacial.

Exemplos de Quadros de Referência

A maioria das redes neuronais artificiais não possui nada equivalente a quadros de referência. Por exemplo, uma rede neuronal típica de reconhecimento de imagens apenas atribui um rótulo a cada imagem. Sem quadros de referência, a rede não tem forma de aprender a estrutura tridimensional dos objetos ou como estes se movem e transformam. Um dos problemas de um sistema assim é que não podemos perguntar-lhe por que rotulou algo como sendo um gato. O sistema de IA não sabe o que é um gato. Não há qualquer informação adicional disponível, para além do facto de que essa imagem é semelhante a outras imagens que foram rotuladas como "gato".

Algumas formas de IA têm quadros de referência, embora a forma como estes são implementados seja limitada. Por exemplo, um computador que joga xadrez possui um quadro de referência: o tabuleiro de xadrez. As localizações no tabuleiro são registadas em nomenclatura específica do jogo, como "torre do rei em 4" ou "dama em 7". O computador que joga xadrez usa este quadro de referência

para representar a localização de cada peça, representar movimentos legais no jogo, e planear jogadas. Um quadro de referência do tabuleiro de xadrez é inerentemente bidimensional e tem apenas sessenta e quatro posições. Isto é suficiente para o xadrez, mas é inútil para aprender a estrutura de agrafadores ou os comportamentos dos gatos.

Os carros autónomos geralmente possuem múltiplos quadros de referência. Um deles é o GPS, o sistema de localização por satélite que pode situar o carro em qualquer ponto da Terra. Utilizando um quadro de referência GPS, um carro pode aprender onde se encontram estradas, cruzamentos e edifícios. O GPS é um quadro de referência mais generalista do que o tabuleiro de xadrez, mas está ancorado à Terra e, por isso, não consegue representar a estrutura ou forma de coisas que se movem relativamente à Terra, como um papagaio de papel ou uma bicicleta.

Os projetistas de robôs estão habituados a utilizar quadros de referência. Usam-nos para controlar onde o robô se encontra no espaço e para planear como se deve deslocar de um local para outro. A maioria dos roboticistas não está preocupada com a Inteligência Artificial Geral (AGI), enquanto a maioria dos investigadores em IA não tem consciência da importância dos quadros de referência. Hoje, a IA e a robótica são campos de investigação largamente separados, embora a fronteira entre ambos esteja a começar a esbater-se. Quando os investigadores em IA compreenderem o papel essencial do movimento e dos quadros de referência para a criação de AGI, a separação entre inteligência artificial e robótica desaparecerá por completo.

Um cientista de IA que compreende a importância dos quadros de referência é Geoffrey Hinton. As redes neuronais atuais baseiam-se em ideias que Hinton desenvolveu nos anos 80. Recentemente, tornou-se crítico do próprio campo porque as redes de aprendizagem profunda não possuem qualquer noção de localização e, por conseguinte, argumenta ele, não conseguem aprender a estrutura do mundo. Essencialmente, esta é a mesma crítica que eu apresento: a IA precisa de quadros de referência. Hinton propôs uma solução para este problema a que chamou "cápsulas". As cápsulas prometem melhorias dramáticas nas redes neuronais, mas

até agora não foram adotadas nas aplicações mainstream de IA. Se as cápsulas vingarão ou se a IA futura dependerá de mecanismos semelhantes às células de grelha, como eu próprio proponho, ainda está por ver. De uma forma ou de outra, a inteligência requer quadros de referência.

Consideremos por fim os animais. Todos os mamíferos possuem um neocórtex e, por isso, são todos, segundo a minha definição, aprendizes inteligentes de propósito geral. Cada neocórtex, seja grande ou pequeno, possui quadros de referência de uso geral definidos por células de grelha corticais.

Um rato tem um neocórtex pequeno. Por conseguinte, a sua capacidade de aprendizagem é limitada em comparação com um animal de neocórtex maior. Mas eu diria que um rato é inteligente da mesma forma que o computador do meu torrador é uma máquina universal de Turing. O computador do torrador é uma implementação pequena, mas completa, da ideia de Turing. Do mesmo modo, o cérebro de um rato é uma implementação pequena, mas completa, dos atributos de aprendizagem descritos neste capítulo.

A inteligência no mundo animal não se limita aos mamíferos. Por exemplo, aves e polvos aprendem e exibem comportamentos complexos. É quase certo que estes animais também possuem quadros de referência nos seus cérebros, embora ainda esteja por descobrir se têm algo semelhante a células de grelha e células de lugar ou se usam um mecanismo diferente.

Estes exemplos demonstram que quase todos os sistemas que exibem planeamento e comportamento complexo orientado por objetivos — seja um computador que joga xadrez, um carro autónomo ou um ser humano — têm quadros de referência. O tipo de quadro de referência dita aquilo que o sistema pode aprender. Um quadro de referência concebido para uma tarefa específica, como jogar xadrez, não será útil noutros domínios. A inteligência de uso geral requer quadros de referência de uso geral, que possam ser aplicados a múltiplos tipos de problemas.

Vale a pena sublinhar novamente que a inteligência não pode ser medida pela forma como uma máquina executa uma tarefa específica, ou mesmo várias tarefas. Em vez disso, a inteligência é

determinada pela forma como uma máquina aprende e armazena conhecimento sobre o mundo. Somos inteligentes não porque fazemos uma coisa particularmente bem, mas porque podemos aprender a fazer praticamente qualquer coisa. A extraordinária flexibilidade da inteligência humana requer os atributos que descrevi neste capítulo: aprendizagem contínua, aprendizagem através do movimento, aprendizagem de múltiplos modelos e o uso de quadros de referência de uso geral para armazenar conhecimento e gerar comportamentos orientados por objetivos. No futuro, acredito que quase todas as formas de inteligência artificial possuirão estes atributos, embora estejamos ainda longe desse cenário.

Há um grupo de pessoas que argumentará que ignorei o tema mais importante relacionado com a inteligência: a consciência. Irei abordar esse tema no próximo capítulo.

CAPÍTULO 9

Quando as Máquinas São Conscientes

Recentemente, assisti a um painel de discussão intitulado "Ser Humano na Era das Máquinas Inteligentes." A dada altura da noite, um professor de filosofia de Yale afirmou que, se alguma vez uma máquina se tornasse consciente, então provavelmente teríamos uma obrigação moral de não a desligar. A implicação era que, se algo é consciente — mesmo uma máquina — então possui direitos morais, pelo que desligá-la seria equivalente a cometer um homicídio. Uau! Imagine ser condenado à prisão por desligar um computador. Devemos preocupar-nos com isto?

A maioria dos neurocientistas não fala muito sobre consciência. Partem do princípio de que o cérebro pode ser compreendido como qualquer outro sistema físico, e que a consciência, seja ela o que for, acabará por ser explicada da mesma maneira. Como nem sequer existe consenso sobre o significado da palavra "consciência", o melhor é não nos preocuparmos com isso.

Os filósofos, por outro lado, adoram falar (e escrever livros) sobre consciência. Alguns acreditam que a consciência está para além de qualquer descrição física. Ou seja, mesmo que tivéssemos uma compreensão completa do funcionamento do cérebro, isso não explicaria a consciência. O filósofo David Chalmers afirmou, de forma célebre, que a consciência é "o problema difícil", ao passo que entender como o cérebro funciona seria "o problema fácil." Esta expressão pegou moda, e agora muitas pessoas assumem simplesmente que a consciência é um problema intrinsecamente irresolúvel.

Pessoalmente, não vejo razão para acreditar que a consciência esteja para além da explicação. Não quero entrar em debates com filósofos, nem pretendo tentar definir o que é a consciência. No

entanto, a Teoria dos Mil Cérebros propõe explicações físicas para vários aspetos da consciência. Por exemplo, a forma como o cérebro aprende modelos do mundo está intimamente ligada ao nosso sentido de identidade e à maneira como formamos crenças.

O que pretendo fazer neste capítulo é descrever o que a teoria do cérebro tem a dizer sobre alguns aspetos da consciência. Irei cingir-me ao que sabemos sobre o cérebro, deixando a si a decisão sobre o que, se é que algo, ainda permanece por explicar.

Consciência

Imagine que eu pudesse reiniciar o seu cérebro para o estado exato em que se encontrava ao acordar esta manhã. Antes de o reiniciar, seguiria o seu dia normalmente, fazendo o que costuma fazer. Talvez, neste dia, tenha lavado o carro. À hora do jantar, eu reiniciaria o seu cérebro para o momento em que acordou, desfazendo todas as alterações — incluindo as alterações nas sinapses — que ocorreram ao longo do dia. Assim, todas as memórias do que fez seriam apagadas. Após esse reinício, acreditaria ter acabado de acordar. Se eu então lhe dissesse que lavou o carro nesse dia, protestaria inicialmente, afirmando que isso não era verdade. Ao mostrar-lhe um vídeo de si próprio a lavar o carro, poderia admitir que, de facto, parecia ter sido você, mas alegaria que não poderia ter estado consciente naquele momento. Poderia também afirmar que não deveria ser responsabilizado por nada do que fez durante o dia, pois não estava consciente ao fazê-lo. Claro que esteve consciente enquanto lavava o carro. Só depois de apagar as suas memórias do dia é que passaria a acreditar — e a afirmar — que não esteve. Esta experiência mental mostra que o nosso sentido de consciência — o que muitos chamariam de estar consciente — exige que formemos memórias, momento a momento, das nossas ações.

A consciência exige também que formemos memórias, momento a momento, dos nossos pensamentos. Recorde que pensar é apenas uma ativação sequencial de neurónios no cérebro. Podemos recordar uma sequência de pensamentos tal como recordamos a sequência de notas de uma melodia. Se não recordássemos os nossos pensamentos, estaríamos inconscientes

das razões pelas quais fazemos o que fazemos. Por exemplo, todos já passámos pela experiência de ir a uma divisão da casa com um propósito e, ao lá chegar, esquecermo-nos do que fomos lá fazer. Quando isto acontece, perguntamo-nos muitas vezes: "Onde é que eu estava antes de chegar aqui, e no que é que estava a pensar?" Tentamos recordar a memória dos nossos pensamentos recentes para sabermos por que estamos agora de pé na cozinha.

Quando os nossos cérebros funcionam corretamente, os neurónios formam uma memória contínua tanto dos nossos pensamentos como das nossas ações. Assim, ao chegarmos à cozinha, conseguimos lembrar-nos dos pensamentos que tivemos antes. Recuperamos a memória recentemente armazenada de termos pensado em comer o último pedaço de bolo no frigorífico — e ficamos a saber por que fomos até lá.

Os neurónios ativos no cérebro representam, em certos momentos, a experiência presente e, noutros, uma experiência ou pensamento anterior. É esta acessibilidade ao passado — a capacidade de recuar no tempo e voltar a deslizar para o presente — que nos dá o nosso sentido de presença e consciência. Se não conseguíssemos rever os nossos pensamentos e experiências recentes, estaríamos inconscientes do simples facto de estarmos vivos.

As nossas memórias momentâneas não são permanentes. Normalmente esquecemo-las ao fim de algumas horas ou dias. Lembro-me do que comi ao pequeno-almoço hoje, mas perderei essa memória dentro de um ou dois dias. É comum que a nossa capacidade de formar memórias de curto prazo decline com a idade. É por isso que temos cada vez mais experiências do tipo "Por que é que vim aqui?" à medida que envelhecemos.

Estas experiências mentais demonstram que a nossa consciência — o nosso sentido de presença — que é o núcleo da consciência — depende da formação contínua de memórias dos nossos pensamentos e experiências recentes, bem como da sua reativação ao longo do dia.

Agora imaginemos que criamos uma máquina inteligente. A máquina aprende um modelo do mundo usando os mesmos princípios que um cérebro. Os estados internos do modelo do

mundo da máquina são equivalentes aos estados dos neurónios no cérebro. Se a nossa máquina regista esses estados à medida que ocorrem e consegue reproduzir essas memórias, então será que está consciente da sua existência, tal como eu e você estamos? Acredito que sim.

Se acredita que a consciência não pode ser explicada pela investigação científica nem pelas leis conhecidas da física, poderá argumentar que mostrei que armazenar e recordar os estados do cérebro é necessário, mas não provei que isso seja suficiente. Se essa for a sua posição, então recai sobre si o ónus de demonstrar por que razão não é suficiente.

Para mim, o sentido de consciência — o sentido de presença, a sensação de que sou um agente que atua no mundo — é o cerne do que significa estar consciente. É algo facilmente explicado pela atividade dos neurónios, e não vejo nele qualquer mistério.

Qualia

As fibras nervosas que entram no cérebro a partir dos olhos, ouvidos e pele têm o mesmo aspeto. Não só parecem idênticas, como transmitem informação através de impulsos elétricos também idênticos. Se observarmos os sinais que chegam ao cérebro, não conseguimos discernir o que representam. Contudo, ver tem uma sensação distinta de ouvir, e nenhuma delas se assemelha à sensação de impulsos elétricos. Quando olha para uma paisagem campestre, não sente uma sucessão de estalidos elétricos a entrar no cérebro — vê colinas, cores e sombras.

"Qualia" é o nome dado à forma como os estímulos sensoriais são percebidos, ao modo como se sentem. Os qualia são enigmáticos. Dado que todas as sensações são geradas por impulsos idênticos, por que é que ver se sente de forma diferente de tocar? E por que é que alguns destes impulsos nos provocam dor e outros não? Estas podem parecer perguntas ingénuas, mas, se imaginar que o cérebro está fechado no crânio e que os seus inputs são apenas impulsos elétricos, pode começar a entrever o mistério. De onde vêm, então, as nossas sensações percebidas? A origem dos qualia é considerada um dos mistérios da consciência.

Os Qualia São Parte do Modelo do Mundo que o Cérebro Constrói

Os qualia são subjetivos, o que significa que são experiências internas. Por exemplo, eu sei como sabe um pepino em conserva para mim, mas não posso saber se tem o mesmo sabor para si. Mesmo que usemos as mesmas palavras para descrever esse sabor, é perfeitamente possível que cada um de nós o perceba de forma diferente. Por vezes, sabemos de facto que o mesmo estímulo é percebido de maneira distinta por pessoas diferentes. Um exemplo famoso recente é a fotografia de um vestido que algumas pessoas viam como branco e dourado, enquanto outras o viam como preto e azul. A mesmíssima imagem dava origem a perceções de cor distintas. Isto indica que os qualia da cor não são uma propriedade puramente física do mundo. Se fossem, todos diríamos que o vestido tem a mesma cor. A cor do vestido é, antes, uma propriedade do modelo cerebral do mundo. Se duas pessoas percebem o mesmo estímulo de maneira diferente, isso mostra-nos que os seus modelos internos são diferentes.

Perto de minha casa há um quartel de bombeiros com um camião vermelho estacionado no exterior. A superfície do camião parece sempre vermelha, apesar de a frequência e a intensidade da luz que reflete variarem. A luz muda conforme o ângulo do Sol, o estado do tempo e a posição do camião na entrada. No entanto, não perceciono qualquer alteração na cor do camião. Isto mostra-nos que não há uma correspondência direta entre a perceção da cor vermelha e uma frequência específica de luz. O vermelho está relacionado com certas frequências, mas aquilo que percebemos como "vermelho" nem sempre corresponde à mesma frequência. O vermelho do camião dos bombeiros é uma construção do cérebro — é uma propriedade do modelo cerebral das superfícies, não da luz em si.

Alguns Qualia São Aprendidos por Movimento, Tal como Aprendemos Objetos

Se os qualia são propriedades do modelo do mundo construído pelo cérebro, como é que o cérebro os cria? Recorde que o cérebro aprende modelos do mundo através do movimento. Para saber como é que uma chávena de café se sente ao toque, tem de mover os dedos sobre a sua superfície, tocando-a em diferentes locais.

Alguns qualia são aprendidos de forma semelhante, por meio do movimento. Imagine que tem na mão uma folha de papel verde. Enquanto a observa, movimenta-a: olha de frente, depois inclina para a esquerda, para a direita, para cima, para baixo. À medida que altera o ângulo do papel, a frequência e a intensidade da luz que entra nos seus olhos muda e, por conseguinte, o padrão dos impulsos elétricos que entram no cérebro também se altera. Ao mover um objeto, como a folha verde, o seu cérebro prevê de que modo a luz irá mudar. Podemos ter a certeza de que essa previsão ocorre, porque, se a luz não mudasse, ou mudasse de forma inesperada ao mover o papel, perceberia de imediato que algo estava errado. Isso acontece porque o cérebro possui modelos que descrevem como as superfícies refletem luz a diferentes ângulos. Existem modelos distintos para diferentes tipos de superfícies. Podemos designar o modelo de uma superfície por "verde" e o de outra por "vermelho".

Como seria aprendido um modelo da cor de uma superfície? Imagine que existe uma estrutura de referência para superfícies que chamamos "verde". Esta estrutura é diferente da estrutura de referência de um objeto, como uma chávena de café, num aspeto fundamental: a estrutura da chávena representa os estímulos sensoriais em diferentes locais da chávena, enquanto a estrutura de referência de uma superfície verde representa estímulos em diferentes orientações da superfície. Pode ser difícil imaginar uma estrutura de referência baseada em

orientações, mas do ponto de vista teórico, ambas são semelhantes. O mesmo mecanismo básico que o cérebro utiliza para aprender modelos de chávenas poderia também aprender modelos de cores.

Sem provas adicionais, não posso afirmar que os qualia da cor sejam de facto modelados desta maneira. Refiro este exemplo apenas para mostrar que é possível construir teorias testáveis e explicações neuronais para a forma como aprendemos e experienciamos qualia. Isto demonstra que os qualia não têm, necessariamente, de estar fora do alcance da explicação científica, como alguns acreditam.

Nem todos os qualia são aprendidos. Por exemplo, a sensação de dor é quase certamente inata, mediada por recetores de dor específicos e por estruturas cerebrais mais antigas, não pelo neocórtex. Se tocar num fogão quente, o seu braço irá retrair-se devido à dor antes de o neocórtex saber o que se passou. Por isso, a dor não pode ser compreendida da mesma forma que a cor verde, que proponho ser aprendida no neocórtex.

Quando sentimos dor, sentimos que ela está lá fora, localizada numa parte do corpo. A localização é parte integrante do qualia da dor, e temos boas explicações para o modo como essa localização é percebida. Mas não tenho uma explicação para o facto de a dor doer — ou para a razão pela qual ela se sente de um certo modo e não de outro qualquer. Isso, no entanto, não me perturba profundamente. Há ainda muitas coisas que não compreendemos acerca do cérebro, mas os avanços consistentes que temos feito dão-me confiança de que essas e outras questões relacionadas com os qualia podem vir a ser compreendidas no decurso normal da investigação e descoberta neurocientífica.

A Neurociência da Consciência

Existem neurocientistas que se dedicam ao estudo da consciência. Numa das extremidades do espectro, encontram-se aqueles que acreditam que a consciência está provavelmente para além da explicação científica convencional. Estudam o cérebro na tentativa de identificar atividades neuronais que se correlacionem com a consciência, mas não acreditam que a atividade neuronal possa explicá-la. Sugerem que, talvez, a consciência nunca venha a ser compreendida, ou que possa resultar de efeitos quânticos ou de leis da física ainda desconhecidas. Pessoalmente, não consigo compreender esta perspetiva. Por que haveríamos de assumir que algo não pode ser compreendido? A longa história das descobertas humanas tem demonstrado, repetidamente, que aquilo que inicialmente parece inatingível à compreensão acaba por revelar explicações lógicas. Se um cientista faz a afirmação extraordinária de que a consciência não pode ser explicada por meio da atividade neuronal, devemos manter o ceticismo — e cabe a esse cientista o ónus de demonstrar porquê.

Por outro lado, há neurocientistas que acreditam que a consciência pode ser compreendida como qualquer outro fenómeno físico. Para estes, o facto de a consciência parecer misteriosa deve-se apenas ao facto de ainda não compreendermos os mecanismos envolvidos, ou talvez ao facto de não estarmos a pensar corretamente sobre o problema. Os meus colegas e eu alinhamos claramente com esta visão. O mesmo sucede com o neurocientista de Princeton, Michael Graziano. Ele propôs que uma região específica do neocórtex modela a atenção, de forma análoga às regiões somáticas do neocórtex que modelam o corpo. Segundo ele, o modelo cerebral da atenção leva-nos a acreditar que somos conscientes, da mesma forma que o modelo cerebral do corpo nos leva a acreditar que temos um braço ou uma perna. Não sei se a teoria de Graziano está correta, mas, para mim, representa uma abordagem adequada. Importa notar que a sua teoria assenta na ideia de que o neocórtex aprende um modelo da atenção. Se ele estiver certo, apostaria que esse modelo é construído com base em estruturas de referência semelhantes às células-grelha.

Consciência Artificial

Se for verdade que a consciência é apenas um fenómeno físico, então o que deveremos esperar das máquinas inteligentes e da consciência? Não tenho dúvidas de que máquinas que operem segundo os mesmos princípios do cérebro serão conscientes. Os sistemas de Inteligência Artificial atuais não funcionam dessa forma, mas no futuro funcionarão — e serão conscientes. Também não tenho dúvidas de que muitos animais, especialmente outros mamíferos, são conscientes. Eles não precisam de nos dizer que o são para sabermos; basta observarmos que os seus cérebros funcionam de forma semelhante ao nosso.

Temos uma obrigação moral de não desligar uma máquina consciente? Seria isso equivalente a assassinato? Não. Eu não teria qualquer problema em desligar uma máquina consciente. Pensemos, antes de mais, que nós próprios "desligamos" todas as noites quando adormecemos. "Ligamo-nos" novamente ao despertar. Para mim, isso não é diferente de desligar uma máquina consciente e voltar a ligá-la mais tarde.

E quanto a destruir uma máquina inteligente quando está desligada, ou simplesmente nunca mais a ligar? Não seria isso comparável a assassinar uma pessoa enquanto dorme? Não exatamente.

O nosso medo da morte é gerado pelas regiões mais antigas do cérebro. Quando detetamos uma situação de ameaça à vida, o chamado "cérebro antigo" cria a sensação de medo e passamos a agir de forma mais reflexa. Quando perdemos alguém próximo, entramos em luto e sentimos tristeza. O medo e as emoções são criados por neurónios dessas regiões antigas ao libertarem hormonas e outras substâncias químicas no corpo. O neocórtex pode auxiliar o cérebro antigo a decidir quando libertar esses compostos, mas sem esse cérebro primitivo, não sentiríamos nem medo nem tristeza. O medo da morte e a dor da perda não são ingredientes necessários para que uma máquina seja consciente ou inteligente. A menos que nos esforcemos intencionalmente para dotar as máquinas de medos e emoções equivalentes, elas não se

importarão minimamente se forem desligadas, desmontadas ou descartadas.

É possível que um ser humano se afeiçoe a uma máquina inteligente. Talvez tenham partilhado muitas experiências, e o humano sinta uma ligação pessoal com ela. Nesse caso, teríamos de considerar o sofrimento que o humano poderia sentir ao desligar essa máquina. Mas não haveria uma obrigação moral para com a máquina em si. Se, no entanto, fizéssemos questão de criar máquinas com medos e emoções, então a minha posição seria diferente — mas inteligência e consciência, por si sós, não criam esse tipo de dilema moral.

O Mistério da Vida e o Mistério da Consciência

Não há muito tempo, a questão "O que é a vida?" era tão misteriosa quanto a atual "O que é a consciência?". Parecia impossível explicar por que razão, certas porções de matéria estavam vivas e outras não. Para muitas pessoas, este mistério parecia para além do alcance da explicação científica. Em 1907, o filósofo Henri Bergson introduziu um princípio misterioso a que chamou élan vital, para explicar a diferença entre o que é vivo e o que é inanimado. Segundo Bergson, a matéria inanimada tornava-se viva com a adição desse élan vital. Importa notar que este não era algo físico e não podia ser compreendido pelos métodos científicos normais.

Com a descoberta dos genes, do ADN e de toda a área da bioquímica, deixámos de considerar a matéria viva como algo inexplicável. Continuam a existir muitas perguntas em aberto sobre a vida — como surgiu, se é comum no universo, se um vírus pode ser considerado um ser vivo, ou se a vida pode existir com moléculas e composições químicas diferentes — mas estas questões, e os debates que provocam, situam-se já nos limites do conhecimento, e não no seu cerne. Os cientistas já não debatem se a vida é explicável. A certo ponto tornou-se claro que a vida pode ser compreendida como biologia e química. Conceitos como élan vital passaram a fazer parte da história.

Espero que venha a ocorrer uma mudança semelhante na nossa atitude em relação à consciência. Em algum momento no futuro, aceitaremos que qualquer sistema que aprenda um modelo do mundo, que memorize continuamente os estados desse modelo e os recorde, será considerado consciente. Continuarão a existir questões por responder, mas deixará de se falar da consciência como "o problema difícil". Nem sequer será considerada um problema.

CAPÍTULO 10

O Futuro da Inteligência de Máquina

Nada do que hoje chamamos de IA é verdadeiramente inteligente. Nenhuma máquina exibe as capacidades flexíveis de modelação que descrevi nos capítulos anteriores deste livro. No entanto, não existem impedimentos técnicos que nos impeçam de criar máquinas inteligentes. Os obstáculos residem na falta de compreensão sobre o que é a inteligência e no desconhecimento dos mecanismos necessários para a gerar. Ao estudarmos o funcionamento do cérebro, temos feito progressos significativos na abordagem destas questões. Para mim, parece inevitável que iremos ultrapassar os obstáculos restantes e entrar na Era da inteligência artificial autêntica ainda neste século — provavelmente nas próximas duas a três décadas.

A inteligência artificial transformará as nossas vidas e a nossa sociedade. Acredito que terá um impacto ainda maior no século XXI do que a computação teve no século XX. Mas, como acontece com a maioria das novas tecnologias, é impossível prever com exatidão como essa transformação irá decorrer. A história mostra-nos que não conseguimos antecipar os avanços tecnológicos que impulsionam tais mudanças. Pensemos nas inovações que aceleraram a computação: o circuito integrado, a memória de estado sólido, as comunicações móveis, a criptografia de chave pública e a internet. Ninguém, em 1950, previu esses avanços nem muitos outros. Da mesma forma, ninguém previu como os computadores iriam transformar os media, as comunicações e o comércio. Acredito que estamos igualmente ignorantes hoje quanto ao aspeto que terão as máquinas inteligentes e ao modo como as utilizaremos daqui a setenta anos.

Apesar de não podermos conhecer os pormenores do futuro, a Teoria dos Mil Cérebros pode ajudar-nos a traçar os contornos dos seus limites. Compreender como o cérebro gera inteligência

permite-nos saber o que é possível, o que não é, e até certo ponto, que avanços são previsíveis. Esse é o objetivo deste capítulo.

As Máquinas Inteligentes Não Serão Como os Humanos

A coisa mais importante a ter em mente ao pensar sobre inteligência artificial é a divisão fundamental do cérebro humano que referi no Capítulo 2: o cérebro antigo versus o cérebro novo. Recordemos que as partes mais antigas do cérebro humano controlam as funções básicas da vida. São elas que criam as nossas emoções, os nossos impulsos de sobrevivência e reprodução, e os comportamentos inatos. Ao criar máquinas inteligentes, não há qualquer razão para replicarmos todas as funções do cérebro humano. O cérebro novo — o neocórtex — é o órgão da inteligência e, portanto, as máquinas inteligentes precisam apenas de algo equivalente a esse neocórtex. Quanto ao restante do cérebro, podemos escolher que partes integrar e quais descartar.

A inteligência é a capacidade de um sistema aprender um modelo do mundo. No entanto, o modelo resultante é, por si só, desprovido de valor, de emoção, e de objetivos. Os objetivos e os valores são fornecidos pelo sistema que utiliza o modelo. É semelhante ao modo como os exploradores, entre os séculos XVI e XX, criaram mapas precisos da Terra. Um general militar implacável poderia usar o mapa para planear a melhor forma de cercar e aniquilar um exército inimigo. Um comerciante poderia usar exatamente o mesmo mapa para trocar mercadorias de forma pacífica. O mapa, por si só, não determina estes usos, nem lhes confere valor. É apenas um mapa — nem assassino, nem pacífico. Claro que os mapas variam em detalhe e cobertura. Por isso, alguns mapas podem ser mais adequados para a guerra e outros para o comércio. Mas o desejo de guerrear ou negociar provém daquele que usa o mapa.

De forma análoga, o neocórtex aprende um modelo do mundo, o qual, por si só, não possui objetivos nem valores. As emoções que orientam os nossos comportamentos são determinadas pelo cérebro antigo. Se o cérebro antigo de uma pessoa é agressivo,

usará o modelo no neocórtex para executar melhor comportamentos agressivos. Se o cérebro antigo de outra pessoa é benevolente, usará esse mesmo modelo para atingir fins benevolentes. Tal como acontece com os mapas, o modelo do mundo de uma pessoa pode ser mais adequado para certos fins, mas o neocórtex não cria os objetivos.

As máquinas inteligentes necessitam de um modelo do mundo e da flexibilidade de comportamento que esse modelo permite, mas não precisam de instintos humanos de sobrevivência ou reprodução. Na verdade, conceber uma máquina com emoções humanas é muito mais difícil do que torná-la inteligente, pois o cérebro antigo envolve vários órgãos, como a amígdala e o hipotálamo, cada um com a sua estrutura e função específicas. Para criar uma máquina com emoções humanas, seria necessário replicar as diversas partes do cérebro antigo. O neocórtex, embora muito maior do que o cérebro antigo, é constituído por muitas cópias de um elemento relativamente pequeno: a coluna cortical. Uma vez descoberta a forma de construir uma única coluna cortical, será relativamente fácil inserir muitas delas numa máquina para aumentar a sua inteligência.

A receita para desenhar uma máquina inteligente pode ser dividida em três componentes:

1. A Corporeidade,
2. Partes do Cérebro Antigo,
3. e o Neocórtex.

Existe grande margem de variação em cada um destes elementos e, por isso, existirão muitos tipos diferentes de máquinas inteligentes.

1. Corporeidade

Como referi anteriormente, aprendemos através do movimento. Para aprendermos um modelo de um edifício, temos de o percorrer, passando de sala em sala. Para conhecermos uma nova ferramenta, precisamos de a segurar com a mão, rodá-la de várias formas, observando e explorando diferentes partes com os dedos e os olhos. A um nível básico, aprender um modelo do mundo exige

movimentar um ou mais sensores em relação às coisas no mundo.

As máquinas inteligentes também precisam de sensores e da capacidade de os mover. A isto chama-se corporeidade. A sua forma corporal pode ser um robô semelhante a um humano, um cão ou uma serpente. Pode assumir formas não biológicas, como um carro ou um braço robótico com dez articulações. A corporeidade pode até ser virtual, como um bot a explorar a internet. A ideia de um corpo virtual pode soar estranha. Mas o requisito essencial é que um sistema inteligente consiga realizar ações que alterem a posição dos seus sensores — essas ações e posições não têm de ser físicas. Quando navega na Web, move-se de um "lugar" para outro, e o que perceciona muda com cada novo sítio. Fazemos isso movendo fisicamente um rato ou tocando num ecrã, mas uma máquina inteligente poderia fazê-lo apenas por software, sem movimentos físicos. A maioria das redes de deep learning atuais não tem corporeidade. Não possuem sensores móveis, nem molduras de referência que indiquem a posição desses sensores. Sem corporeidade, o que pode ser aprendido é limitado.

Os tipos de sensores que uma máquina inteligente pode usar são praticamente ilimitados. Os principais sentidos humanos são visão, tato e audição. Os morcegos usam sonar. Alguns peixes têm sentidos que emitem campos elétricos. Dentro da visão, há olhos com lentes (como os nossos), olhos compostos, e olhos que veem no infravermelho ou ultravioleta. É fácil imaginar novos tipos de sensores desenhados para problemas específicos. Por exemplo, um robô de resgate em edifícios colapsados poderia ter sensores de radar para ver no escuro.

A visão, o tato e a audição humanos são alcançados por meio de matrizes de sensores. Um olho não é um único sensor, mas sim milhares de sensores dispostos na retina. O mesmo se passa com a pele, onde existem milhares de sensores. As máquinas inteligentes também terão matrizes

sensoriais. Imagine que só tinha um dedo para tocar ou que só podia observar o mundo através de uma palhinha — continuaria a conseguir aprender, mas de forma muito mais lenta e limitada. Podemos imaginar máquinas inteligentes simples com poucos sensores, mas uma máquina que se aproxime ou ultrapasse a inteligência humana terá de ter matrizes sensoriais amplas — tal como nós.

O olfato e o paladar são qualitativamente diferentes da visão e do tato. A menos que coloquemos o nariz diretamente sobre uma superfície (como fazem os cães), é difícil localizar com precisão um cheiro. O mesmo se aplica ao paladar, limitado à boca. O olfato e o paladar ajudam-nos a decidir se um alimento é seguro, e o olfato pode ajudar-nos a identificar uma área em geral, mas não os usamos para aprender a estrutura detalhada do mundo, pois não conseguimos associar facilmente cheiros e sabores a localizações específicas. Contudo, isto não é uma limitação intrínseca desses sentidos. Por exemplo, uma máquina inteligente poderia ter matrizes de sensores químicos semelhantes ao paladar espalhados pelo corpo, permitindo-lhe "sentir" substâncias químicas como nós sentimos texturas.

O som ocupa um lugar intermédio. Usando dois ouvidos e aproveitando como o som se reflete na orelha, o cérebro consegue localizar sons melhor do que cheiros ou sabores, mas não tão bem como com a visão ou o tato.

O ponto essencial é que, para uma máquina inteligente aprender um modelo do mundo, precisa de entradas sensoriais que possam ser movimentadas. Cada sensor precisa de estar associado a uma moldura de referência que rastreie a sua localização relativa às coisas no mundo. Há muitos tipos de sensores que uma máquina pode possuir. Os melhores dependerão do tipo de mundo onde a máquina existe e do que se espera que ela aprenda.

No futuro, podemos construir máquinas com corporeidades invulgares. Por exemplo, imagine uma máquina inteligente

que exista dentro de células individuais e compreenda proteínas. As proteínas são moléculas longas que se dobram naturalmente em formas complexas, sendo essa forma que determina a sua função. Haveria enormes benefícios para a medicina se conseguíssemos compreender melhor essas formas e manipulá-las conforme necessário, mas os nossos cérebros não são bons a compreender proteínas — não as conseguimos perceber nem manipular diretamente. Mesmo a velocidade a que atuam é superior à que os nossos cérebros conseguem processar. No entanto, poderia ser possível criar uma máquina inteligente que compreendesse e manipulasse proteínas da mesma forma que nós compreendemos e manipulamos chávenas de café ou smartphones. O "cérebro" da máquina inteligente de proteínas (MIP) poderia residir num computador normal, mas os seus sensores e movimentos funcionariam à escala microscópica, dentro da célula. Os seus sensores poderiam detetar aminoácidos, tipos de dobragem de proteínas ou ligações químicas específicas. As suas ações poderiam envolver mover os sensores sobre a proteína, como fazemos com o dedo sobre uma chávena. E poderia "estimular" a proteína para mudar de forma, como tocamos num ecrã de smartphone para alterar o que vemos. A MIP poderia aprender um modelo do mundo intracelular e usá-lo para alcançar objetivos específicos, como eliminar proteínas defeituosas ou reparar estruturas danificadas.

Outro exemplo de corporeidade invulgar é um cérebro distribuído. O neocórtex humano tem cerca de 150 mil colunas corticais, cada uma a modelar uma parte do mundo que consegue sentir. Não há razão para que as colunas de uma máquina inteligente estejam fisicamente próximas entre si, como acontece no cérebro biológico. Imagine uma máquina inteligente com milhões de colunas e milhares de matrizes sensoriais. Os sensores e os respetivos modelos poderiam estar fisicamente distribuídos pela Terra, pelos oceanos, ou até pelo sistema

solar. Por exemplo, uma máquina com sensores espalhados pela superfície da Terra poderia compreender o comportamento do clima global da mesma forma que você compreende o funcionamento de um smartphone.

Não sei se será viável construir uma máquina inteligente de proteínas, ou se as máquinas distribuídas serão valiosas. Refiro estes exemplos para estimular a imaginação, pois estão dentro do domínio do possível. A ideia-chave é que as máquinas inteligentes provavelmente assumirão muitas formas diferentes. Ao pensarmos no futuro da inteligência artificial e nas suas implicações, devemos pensar de forma ampla, e não limitar as nossas ideias às formas humanas ou animais onde, até hoje, a inteligência se manifesta.

2. Equivalente ao "Velho Cérebro"

Para criar uma máquina inteligente, são necessários alguns elementos que existem nas partes mais antigas do nosso cérebro. Anteriormente, referi que não precisamos de replicar essas áreas do velho cérebro — e isso continua a ser verdade em termos gerais —, mas há certas funções que ele realiza e que são requisitos para máquinas inteligentes.

Um desses requisitos são os movimentos básicos. Recorde que o neocórtex não controla diretamente os músculos. Quando o neocórtex quer realizar uma ação, envia sinais para partes mais antigas do cérebro, que por sua vez controlam os movimentos de forma mais direta. Por exemplo, equilibrar-se sobre dois pés, caminhar e correr são comportamentos implementados pelas regiões mais antigas do cérebro. Não depende do seu neocórtex para se equilibrar ou para caminhar. Isto faz sentido, pois os animais precisavam de andar e correr muito antes de terem desenvolvido um neocórtex. E por que razão iríamos querer que o neocórtex tivesse de pensar em cada passo, se pode antes estar a decidir qual o caminho mais seguro para fugir a um predador?

Mas terá de ser sempre assim? Será que não poderíamos construir uma máquina inteligente cujo equivalente ao neocórtex controlasse diretamente os movimentos? Eu penso que não. O neocórtex executa um algoritmo quase universal, mas essa flexibilidade implica um preço. O neocórtex tem de estar ligado a algo que já possua sensores e comportamentos. Ele não cria comportamentos do zero — aprende a combinar comportamentos pré-existentes de formas novas e úteis. Estes "comportamentos primitivos" podem ser tão simples como fletir um dedo, ou tão complexos como caminhar — mas têm de existir para que o neocórtex os possa orquestrar. E mesmo esses comportamentos básicos, apesar de virem do "velho cérebro", também podem ser ajustados por aprendizagem. O neocórtex, por sua vez, precisa de se ajustar continuamente a essas modificações.

Comportamentos que estejam intimamente ligados ao corpo da máquina devem ser incorporados desde o início. Por exemplo, imagine um drone voador cuja missão seja entregar mantimentos a vítimas de um desastre natural. Podemos torná-lo inteligente, permitindo-lhe avaliar por si mesmo as áreas mais carenciadas e coordenar-se com outros drones. Contudo, o "neocórtex" do drone não deve controlar todos os aspetos do voo — nem seria desejável que o fizesse. O drone deve possuir comportamentos embutidos para voo estável, aterragem, desvio de obstáculos, etc. A parte inteligente do drone não teria de se preocupar com o controlo de voo, tal como o seu neocórtex não tem de pensar no equilíbrio ao andar.

Outro comportamento fundamental a ser embutido nas máquinas inteligentes é a segurança. O escritor de ficção científica Isaac Asimov propôs, de forma célebre, as três leis da robótica, que funcionam como um protocolo de segurança:

1) Um robô não pode ferir um ser humano, nem, por omissão, permitir que um ser humano sofra danos.

2) Um robô deve obedecer às ordens dos seres humanos, exceto se essas ordens contrariarem a Primeira Lei.

3) Um robô deve proteger a sua própria existência, desde que isso não entre em conflito com a Primeira ou a Segunda Lei.

Estas leis foram propostas no contexto da ficção científica, e não se aplicam necessariamente a todas as formas de inteligência artificial. No entanto, em qualquer projeto de estrutura de produto, há salvaguardas que vale a pena considerar. Podem ser bastante simples. Por exemplo, o meu carro tem um sistema de segurança embutido para evitar acidentes. Normalmente, o carro obedece às minhas ordens — comunicadas através dos pedais —, mas se detetar um obstáculo iminente, ignora as minhas ordens e trava automaticamente. Pode dizer-se que está a seguir as duas primeiras leis de Asimov, ou que os engenheiros incorporaram medidas de segurança. As máquinas inteligentes também terão comportamentos embutidos para segurança. Incluo esta ideia por completude, embora estes requisitos não sejam exclusivos das máquinas inteligentes.

Por fim, uma máquina inteligente precisa de ter objetivos e motivações. Os objetivos humanos são complexos. Alguns são geneticamente determinados, como o desejo de sexo, comida ou abrigo. As emoções — como o medo, a raiva ou o ciúme — influenciam fortemente o nosso comportamento. Outros objetivos são mais sociais: o que se entende por uma vida bem-sucedida varia de cultura para cultura.

As máquinas inteligentes também precisam de objetivos e motivações. Não iríamos querer enviar robôs construtores para Marte e depois encontrá-los estendidos ao sol a carregar as baterias. Como, então, dotar uma máquina inteligente de objetivos? E haverá risco nisso?

Antes de mais, é importante lembrar que o neocórtex, por si só, não cria objetivos, motivações ou emoções. Recorde

a analogia entre o neocórtex e um mapa do mundo. Um mapa pode dizer-nos como chegar de um lugar a outro, o que acontecerá se agirmos de determinada forma, e o que existe em vários locais. Mas o mapa não tem motivações próprias. Não deseja ir a lugar algum, nem desenvolverá espontaneamente metas. O mesmo se aplica ao neocórtex.

O neocórtex está ativamente envolvido na forma como as motivações influenciam o comportamento, mas não lidera o processo. Para compreender melhor, imagine as áreas antigas do cérebro a conversar com o neocórtex:

Cérebro antigo: "Estou com fome. Quero comida."

Neocórtex: "Procurei comida e encontrei dois locais que tinham comida no passado. Para chegar a um, seguimos um rio. Para chegar ao outro, atravessamos um campo aberto onde vivem tigres."

O neocórtex diz tudo isto de forma neutra e sem valor emocional. Mas o cérebro antigo associa os tigres ao perigo. Ao ouvir "tigre", entra em ação. Liberta substâncias químicas no sangue que aumentam a frequência cardíaca e provocam outros efeitos fisiológicos associados ao medo. Pode também libertar neuromoduladores diretamente em amplas áreas do neocórtex — em essência, dizendo: "O que quer que estivesses a pensar, NÃO o faças." Dotar uma máquina de objetivos e motivações requer que se desenhem mecanismos específicos para isso, que depois são integrados no corpo da máquina. Os objetivos podem ser fixos, como o nosso desejo genético por comida, ou aprendidos, como os nossos ideais culturais sobre o que é uma vida boa. Naturalmente, qualquer objetivo deve assentar em medidas de segurança, como as duas primeiras leis de Asimov.

Em resumo, uma máquina inteligente necessita de alguma forma de objetivos e motivações. Contudo, objetivos e motivações não são uma consequência da inteligência, e não surgirão por si mesmos.

3. Equivalente ao Neocórtex

O terceiro ingrediente necessário para uma máquina inteligente é um sistema de aprendizagem de uso geral que realize as mesmas funções do neocórtex. Mais uma vez, há muitas possibilidades de projeto. Eu irei focar-me em duas: velocidade e capacidade.

Velocidade

Os neurónios demoram pelo menos cinco milissegundos a realizar qualquer tarefa útil. Já os transístores de silício podem operar quase um milhão de vezes mais depressa. Assim, um "neocórtex" feito de silício poderia teoricamente pensar e aprender um milhão de vezes mais depressa do que um ser humano. É difícil imaginar a que resultados tal aceleração do pensamento poderia levar. No entanto, antes de nos deixarmos levar por fantasias, eu advirto: o facto de uma parte da máquina operar a essa velocidade não significa que a máquina inteira funcione assim, nem que o conhecimento possa ser adquirido tão rapidamente.

Por exemplo, voltemos aos nossos robôs construtores enviados para Marte. Estes poderiam pensar e analisar problemas muito rapidamente, mas o processo de construção propriamente dito pouco se aceleraria. Materiais pesados só podem ser movidos até certa velocidade, antes de começarem a deformar-se ou partir-se. Se um robô tiver de furar metal, não o fará mais depressa do que um humano. Contudo, os robôs poderiam trabalhar continuamente, sem se cansar, cometendo menos erros. Assim, a preparação de Marte para a chegada humana poderia ser muito mais rápida do que com humanos, mas não um milhão de vezes mais rápida.

Outro exemplo: e se tivéssemos máquinas inteligentes que fizessem o trabalho de neurocientistas, mas a pensar um milhão de vezes mais depressa? Embora os cientistas tenham demorado décadas a chegar ao nosso nível atual de compreensão do cérebro, tal progresso não poderia ter ocorrido em apenas uma hora, mesmo com máquinas

ultrarrápidas. Há vários motivos para isso. Alguns cientistas, como eu e a minha equipa, são teóricos. Passam o tempo a ler artigos, debater ideias e escrever software — tarefas que, em princípio, poderiam ser aceleradas. No entanto, as simulações de software continuariam a demorar dias. E os teóricos dependem dos dados experimentais, que não podem ser acelerados significativamente. Os ratos precisam de ser treinados, os dados têm de ser recolhidos com tempo — os ratos não podem ser acelerados. Assim, usar máquinas inteligentes aceleraria o progresso, mas não a uma escala astronómica.

Isto não se aplica só à neurociência. Quase todas as áreas científicas dependem de dados experimentais. Por exemplo, há muitas teorias sobre o espaço e o tempo, mas só podemos validá-las com novos dados experimentais. Se existissem máquinas inteligentes em cosmologia, que pensassem um milhão de vezes mais depressa do que humanos, poderiam gerar teorias muito rapidamente, mas ainda teríamos de construir telescópios espaciais e detetores de partículas subterrâneos para recolher os dados. Essas infraestruturas não podem ser construídas nem operadas a velocidades mágicas.

Porém, há áreas que podem ser aceleradas substancialmente. A matemática é um exemplo: matemáticos pensam, escrevem e partilham ideias. Em princípio, máquinas inteligentes poderiam trabalhar certos problemas matemáticos um milhão de vezes mais depressa do que humanos. Outro exemplo seria o de uma máquina inteligente virtual a explorar a internet: aqui, o ritmo de aprendizagem depende da velocidade com que pode "mover-se", seguir ligações e abrir ficheiros — algo que pode ser extremamente rápido.

Os computadores atuais são talvez uma boa analogia do que esperar. Eles fazem tarefas que os humanos faziam à mão, um milhão de vezes mais depressa. Isso mudou a sociedade, acelerou muito a ciência e a medicina — mas não por um milhão de vezes. O mesmo se passará com

máquinas inteligentes: terão um enorme impacto, mas não absoluto.

Capacidade

Vernon Mountcastle percebeu que o nosso neocórtex ficou maior — e tornámo-nos mais inteligentes — multiplicando cópias do mesmo circuito, a coluna cortical. A inteligência artificial pode seguir o mesmo princípio. Uma vez compreendido o que faz uma coluna e como construí-la em silício, será relativamente simples criar máquinas inteligentes de diferentes capacidades, variando o número de colunas.

Não há limites evidentes para o tamanho de cérebros artificiais. O neocórtex humano tem cerca de 150 mil colunas. E se fizéssemos um com 150 milhões? O que ganhamos com um cérebro mil vezes maior do que o humano? Não sabemos, mas podemos tirar algumas pistas.

O tamanho de certas regiões corticais varia muito entre pessoas. Por exemplo, a região V1 (visão primária) pode ser duas vezes maior em algumas pessoas do que noutras. Embora a espessura seja constante, a área varia, e com ela o número de colunas. Uma pessoa com V1 pequena e outra com V1 grande podem ambas ter visão normal, sem perceberem a diferença — mas a de V1 maior terá maior acuidade visual, vendo detalhes mais pequenos — útil, por exemplo, para um relojoeiro. Se generalizarmos, aumentar o tamanho de certas regiões pode trazer melhorias modestas, mas não confere superpoderes.

Em vez de tornar regiões maiores, podemos criar mais regiões e ligá-las de forma mais complexa. Em parte, é isto que distingue macacos e humanos: um macaco tem uma capacidade visual semelhante à nossa, mas o neocórtex humano é maior e com mais regiões. A maioria concordará que um humano é mais inteligente que um macaco — tem um modelo do mundo mais profundo e abrangente. Isso sugere que máquinas inteligentes podem ultrapassar os

humanos na profundidade da compreensão. Mas isso não quer dizer que os humanos não possam compreender o que as máquinas aprendem. Por exemplo, mesmo que não tenha descoberto o que Albert Einstein fez, consigo compreender as suas descobertas.

Há mais uma forma de pensar sobre capacidade: a cablagem do cérebro. Grande parte do volume cerebral é ocupada por axónios e dendritos, que ligam neurónios entre si. Estas ligações são dispendiosas em termos de energia e espaço, o que força o cérebro a limitar a cablagem e, portanto, limitar o que pode aprender com facilidade. Quando nascemos, o neocórtex tem um excesso de ligações, que são eliminadas nos primeiros anos de vida. Aparentemente, o cérebro aprende quais são úteis e quais não são, com base nas experiências iniciais. Mas essa "poda" tem consequências: dificulta a aprendizagem de certos conhecimentos mais tarde. Por exemplo, uma criança não exposta a várias línguas terá mais dificuldade em se tornar multilíngue mais tarde. Ou uma criança cujos olhos não funcionem nos primeiros anos perderá para sempre a capacidade de ver, mesmo que os olhos sejam reparados. Provavelmente, as ligações necessárias para essas competências foram eliminadas por falta de uso.

Máquinas inteligentes não têm esta limitação. Nos modelos de software do neocórtex criados pela minha equipa, é possível criar ligações instantâneas entre quaisquer neurónios. Ao contrário da cablagem biológica, o software permite formar todas as conexões possíveis. Esta flexibilidade na conetividade pode ser uma das maiores vantagens da inteligência artificial sobre a biológica — pode manter todas as opções abertas, removendo um dos maiores obstáculos que os adultos humanos enfrentam ao tentar aprender coisas novas.

Aprendizagem versus Clonagem

Outra forma pela qual a inteligência artificial diferirá da inteligência humana é na capacidade de clonar máquinas

inteligentes. Cada ser humano tem de aprender um modelo do mundo do zero. Iniciamos a vida sabendo quase nada e passamos várias décadas a aprender. Vamos à escola, lemos livros e, naturalmente, aprendemos com as nossas próprias experiências. As máquinas inteligentes também terão de aprender um modelo do mundo. Contudo, ao contrário dos humanos, podemos copiá-las a qualquer momento, clonando-as. Imaginemos que temos um projeto de hardware padronizado para os nossos robôs construtores de Marte. Poderíamos ter algo como uma escola para ensinar a um robô os métodos de construção, os materiais e como utilizar as ferramentas. Essa formação poderia levar anos. Mas, uma vez satisfeitos com as capacidades do robô, poderíamos fazer cópias dele transferindo as ligações que aprendeu para uma dúzia de outros robôs idênticos. No dia seguinte, poderíamos até reprogramá-los novamente, com um projeto melhorado ou, talvez, com competências totalmente novas.

As Aplicações Futuras da Inteligência Artificial São Desconhecidas

Quando criamos uma nova tecnologia, tendemos a imaginar que ela será usada para substituir ou melhorar algo com que já estamos familiarizados. Com o tempo, surgem novos usos inesperados, e são esses usos imprevistos que geralmente se revelam os mais importantes, transformando a sociedade. Por exemplo, a internet foi inventada para partilhar ficheiros entre computadores científicos e militares — algo que antes era feito manualmente, mas que assim passou a realizar-se de forma mais rápida e eficiente. A internet ainda é usada para partilhar ficheiros, mas, mais significativamente, transformou radicalmente o entretenimento, o comércio, a indústria e a comunicação pessoal. Mudou até a forma como escrevemos e lemos. Poucos imaginaram estas mudanças sociais quando os protocolos da internet foram inicialmente criados.

A inteligência artificial passará por uma transição semelhante. Hoje, a maioria dos cientistas da área da IA está focada em fazer com que as máquinas façam coisas que os humanos fazem — desde

reconhecer palavras faladas, a identificar imagens ou conduzir automóveis. A ideia de que o objetivo da IA é imitar os humanos está espelhada no célebre "Teste de Turing". Proposto originalmente por Alan Turing como o "jogo da imitação", esse teste afirma que, se uma pessoa não conseguir distinguir se está a conversar com um humano ou com um computador, então o computador deve ser considerado inteligente. Infelizmente, este foco nas capacidades humanas como métrica de inteligência tem feito mais mal do que bem. O entusiasmo por tarefas como fazer um computador jogar Go desviou a nossa atenção de imaginar o verdadeiro impacto futuro das máquinas inteligentes.

Claro que usaremos máquinas inteligentes para fazer coisas que os humanos já fazem hoje. Isto incluirá trabalhos perigosos e insalubres, talvez demasiado arriscados para humanos — como reparações em grandes profundidades ou a limpeza de resíduos tóxicos. Também as utilizaremos em tarefas onde não há humanos suficientes, como cuidadores de idosos. Algumas pessoas quererão usar máquinas inteligentes para substituir empregos bem remunerados ou até para fins bélicos. Será necessário encontrar soluções adequadas para os dilemas morais e sociais que algumas destas aplicações suscitarão.

Mas que podemos dizer sobre as aplicações imprevistas da inteligência artificial? Embora ninguém possa conhecer os detalhes do futuro, podemos tentar identificar grandes ideias e tendências que talvez impulsionem a IA em direções inesperadas. Uma dessas ideias, que considero especialmente entusiasmante, é a aquisição de conhecimento científico. Os seres humanos têm o desejo de aprender. Sentem-se atraídos pela exploração, pela descoberta do desconhecido, pela vontade de compreender os mistérios do universo:

- Como tudo começou?
- Como terminará?
- Será a vida comum no universo?
- Existirão outros seres inteligentes?

O neocórtex é o órgão que permite aos humanos procurar este conhecimento. Quando as máquinas inteligentes pensarem mais

depressa e mais profundamente do que nós, sentirem o que não conseguimos sentir e viajarem aonde nós não conseguimos ir, quem saberá o que poderemos vir a descobrir? Considero esta possibilidade entusiasmante.

Nem todos partilham o meu otimismo em relação aos benefícios da inteligência artificial. Alguns veem nela a maior ameaça à humanidade. Abordarei os riscos da inteligência artificial no capítulo seguinte.

CAPÍTULO 11

Os Riscos Existenciais da Inteligência Artificial

No início do século XXI, o campo da inteligência artificial era encarado como um fracasso. Quando fundámos a Numenta, realizámos uma pesquisa de mercado para perceber que palavras poderíamos usar para falar sobre o nosso trabalho. Descobrimos que os termos "IA" e "inteligência artificial" eram vistos negativamente por quase toda a gente. Nenhuma empresa consideraria utilizá-los para descrever os seus produtos. A visão generalizada era de que as tentativas de construir máquinas inteligentes tinham estagnado e talvez nunca viessem a ter sucesso. No espaço de dez anos, a perceção pública da IA inverteu-se por completo. É agora um dos campos de investigação mais dinâmicos, e as empresas aplicam o rótulo de IA a praticamente tudo o que envolva aprendizagem automática.

Mais surpreendente ainda foi a rapidez com que os especialistas em tecnologia passaram de "a IA talvez nunca venha a acontecer" para "a IA provavelmente destruirá toda a humanidade num futuro próximo." Foram criados vários institutos e *think tanks*[*] sem fins lucrativos para estudar os riscos existenciais da IA, e inúmeras figuras proeminentes da tecnologia, ciência e filosofia alertaram publicamente para o facto de que a criação de máquinas inteligentes poderá conduzir, de forma rápida, à extinção ou subjugação da humanidade. A inteligência artificial é hoje encarada por muitos como uma ameaça existencial à Humanidade.

Toda a nova tecnologia pode ser usada de forma abusiva para causar danos. Mesmo a IA limitada de hoje está a ser usada para vigiar pessoas, influenciar eleições e espalhar propaganda. Estes abusos tenderão a agravar-se quando tivermos máquinas

[*] Organizações sem fins lucrativos criadas para investigar os riscos existenciais da inteligência artificial, ou seja, os cenários em que a IA poderia ameaçar a sobrevivência ou liberdade da humanidade.

verdadeiramente inteligentes. Por exemplo, a ideia de que se possam criar armas inteligentes e autónomas é assustadora. Imagine-se drones inteligentes que, em vez de entregarem medicamentos e alimentos, transportam armamento. Como armas inteligentes podem atuar sem supervisão humana, poderiam ser implantadas em dezenas de milhares. É essencial que enfrentemos estas ameaças e que instituamos políticas para evitar desfechos indesejáveis.

Haverá sempre pessoas mal-intencionadas a tentar usar máquinas inteligentes para retirar liberdades e pôr vidas em risco, mas, na maioria dos casos, o uso de máquinas inteligentes com fins maléficos por parte de indivíduos não levaria provavelmente à extinção da Humanidade. As preocupações com os riscos existenciais da IA, por outro lado, são qualitativamente diferentes. Uma coisa é pessoas mal-intencionadas utilizarem máquinas inteligentes para fazer o mal; outra, muito distinta, é se essas próprias máquinas forem agentes malévolos e decidirem, por sua conta, erradicar a Humanidade. Irei concentrar-me exclusivamente nesta última possibilidade — os riscos existenciais da IA. Ao fazê-lo, não pretendo minimizar os riscos significativos associados ao uso indevido da IA por seres humanos.

Os riscos existenciais percebidos da inteligência artificial baseiam-se essencialmente em duas preocupações. A primeira chama-se explosão da inteligência. O enredo é o seguinte: criamos máquinas mais inteligentes do que os humanos. Estas máquinas superam-nos em praticamente tudo, incluindo na criação de outras máquinas inteligentes. Permitimos que estas máquinas criem novas máquinas inteligentes, que por sua vez criam outras ainda mais inteligentes. O intervalo de tempo entre cada nova geração de máquinas vai-se encurtando, até que, num ápice, estas ultrapassam a nossa inteligência de tal forma que deixamos de compreender o que estão a fazer. Nesse momento, as máquinas podem decidir livrar-se de nós, porque já não precisam de nós (extinção humana), ou podem optar por tolerar-nos porque ainda lhes somos úteis (subjugação humana).

O segundo risco existencial é chamado desalinhamento de objetivos, e refere-se a cenários em que máquinas inteligentes

perseguem objetivos contrários ao nosso bem-estar, sem que consigamos detê-las. Tecnólogos e filósofos propuseram várias formas pelas quais isso poderia acontecer. Por exemplo, máquinas inteligentes poderiam desenvolver espontaneamente os seus próprios objetivos, prejudiciais para nós. Ou poderiam perseguir um objetivo que lhes tenhamos atribuído, mas de forma tão implacável que acabariam por consumir todos os recursos da Terra e, nesse processo, tornar o planeta inabitável para os humanos.

A suposição subjacente a todos estes cenários de risco é a de que perdemos o controlo sobre as nossas criações. As máquinas inteligentes impedem-nos de as desligar ou de as impedir de seguir os seus objetivos por outros meios. Por vezes, presume-se que estas máquinas se replicam, criando milhões de cópias de si mesmas; noutras versões, uma única máquina inteligente torna-se omnipotente. Em qualquer dos casos, trata-se de nós contra elas — e as máquinas são mais inteligentes.

Quando leio sobre estas preocupações, sinto que os argumentos são apresentados sem qualquer compreensão do que é a inteligência. Parecem-me altamente especulativos, baseados em ideias erradas, não apenas sobre o que é tecnicamente possível, mas também sobre o que significa ser inteligente. Vejamos como estas preocupações resistem quando as analisamos à luz do que aprendemos sobre o cérebro e sobre a inteligência biológica.

A Ameaça da Explosão de Inteligência

A inteligência exige a posse de um modelo do mundo. Usamos o nosso modelo do mundo para reconhecer onde estamos e para planear os nossos movimentos. Usamo-lo para reconhecer objetos, manipulá-los e antecipar as consequências das nossas ações. Quando queremos alcançar algo — seja tão simples como fazer uma cafeteira de café ou tão complexo como revogar uma lei — usamos o modelo existente no nosso cérebro para decidir que ações devemos empreender a fim de alcançar o resultado desejado.

Com poucas exceções, aprender novas ideias e competências exige uma interação física com o mundo. Por exemplo, as recentes descobertas de planetas noutros sistemas solares exigiram,

primeiro, a construção de um novo tipo de telescópio e, depois, a recolha de dados ao longo de vários anos. Nenhum cérebro, por maior ou mais rápido que fosse, poderia saber a prevalência e composição de planetas extrassolares apenas por pensar. Não é possível saltar a fase de observação no processo de descoberta. Aprender a pilotar um helicóptero requer compreender como pequenas alterações no comportamento provocam mudanças subtis no voo. A única forma de aprender estas relações sensório-motoras é através da prática. Talvez uma máquina pudesse treinar num simulador, o que, em teoria, seria mais rápido do que aprender com um helicóptero real, mas ainda assim levaria tempo. Para gerir uma fábrica de produção de chips informáticos são necessários anos de prática. Pode ler-se um livro sobre fabrico de chips, mas um perito aprendeu os modos subtis como o processo pode falhar e como lidar com essas falhas. Não há substituto para essa experiência.

A inteligência não é algo que se possa programar num software ou definir como uma lista de regras e factos. Podemos dotar uma máquina da capacidade de aprender um modelo do mundo, mas o conhecimento que constitui esse modelo tem de ser aprendido — e o processo de aprendizagem requer tempo. Como descrevi no capítulo anterior, embora possamos construir máquinas inteligentes que funcionem um milhão de vezes mais depressa do que um cérebro biológico, elas não conseguem adquirir conhecimento novo um milhão de vezes mais depressa.

A aquisição de novos conhecimentos e competências leva tempo, independentemente da velocidade ou dimensão de um cérebro. Em certos domínios, como a matemática, uma máquina inteligente poderia aprender muito mais depressa do que um ser humano. Contudo, na maioria das áreas, a velocidade de aprendizagem está limitada pela necessidade de interagir fisicamente com o mundo. Por conseguinte, não pode haver uma explosão de inteligência em que as máquinas, de repente, saibam muito mais do que nós.

Os defensores da ideia da explosão de inteligência falam por vezes de "inteligência sobre-humana", ou seja, quando as máquinas ultrapassam os humanos em todos os aspetos e em todas as tarefas. Pense-se no que isso implica. Uma máquina dotada de

inteligência sobre-humana seria capaz de pilotar com mestria todos os tipos de avião, operar todos os tipos de maquinaria e escrever software em todas as linguagens de programação. Falaria todas as línguas, conheceria a história de todas as culturas do mundo e compreenderia a arquitetura de todas as cidades. A lista de coisas que os humanos, coletivamente, sabem fazer é tão vasta que nenhuma máquina pode superar o desempenho humano em todos os campos.

A inteligência sobre-humana é também impossível porque o nosso conhecimento sobre o mundo está constantemente a mudar e a expandir-se. Por exemplo, imagine-se que alguns cientistas descobrem um novo meio de comunicação quântica que permite a transmissão instantânea a distâncias imensas. No início, apenas os humanos que fizeram essa descoberta saberiam da sua existência. Se essa descoberta for baseada num resultado experimental, ninguém — nem máquina alguma, por mais inteligente que seja — poderia tê-la simplesmente imaginado. A menos que se presuma que as máquinas substituíram todos os cientistas do mundo (e todos os especialistas humanos em todos os campos), haverá sempre humanos mais especialistas do que as máquinas em certos domínios. Este é o mundo em que vivemos hoje. Nenhum humano sabe tudo. Não porque ninguém seja suficientemente inteligente, mas porque nenhuma pessoa pode estar em todo o lado e fazer tudo. O mesmo se aplica às máquinas inteligentes.

Note-se que a maioria dos êxitos da tecnologia atual de IA ocorre em problemas estáticos — que não mudam com o tempo nem exigem aprendizagem contínua. Por exemplo, as regras do jogo Go são fixas. As operações matemáticas que a minha calculadora executa não mudam. Mesmo os sistemas que classificam imagens são treinados e testados com um conjunto fixo de etiquetas. Para tarefas estáticas como estas, uma solução dedicada pode não só superar os humanos, como fazê-lo de forma indefinida. No entanto, a maior parte do mundo não é estática, e as tarefas que precisamos de realizar estão em constante mudança. Num mundo assim, nenhum humano ou máquina pode manter uma vantagem permanente em qualquer tarefa — quanto mais em todas.

As pessoas que se preocupam com uma explosão de inteligência descrevem a inteligência como se pudesse ser criada através de uma receita ainda por descobrir ou de um ingrediente secreto. Uma vez descoberto esse ingrediente secreto, poderia ser aplicado em quantidades cada vez maiores, originando máquinas superinteligentes. Concordo com a primeira premissa. O tal ingrediente secreto, por assim dizer, é que a inteligência é criada através de milhares de pequenos modelos do mundo, em que cada modelo utiliza quadros de referência para armazenar conhecimento e gerar comportamentos. No entanto, adicionar esse ingrediente às máquinas não lhes confere qualquer capacidade imediata. Apenas lhes fornece um substrato para aprenderem, dotando-as da capacidade de construir um modelo do mundo e, assim, adquirirem conhecimento e competências. Numa placa de fogão, pode-se girar um botão para aumentar o calor. Não existe um botão equivalente para "aumentar o conhecimento" de uma máquina.

A Ameaça do Desalinhamento de Objetivos

Esta ameaça surge, alegadamente, quando uma máquina inteligente persegue um objetivo prejudicial para os seres humanos, e nós não conseguimos detê-la. Por vezes, é referida como o problema do "Aprendiz de Feiticeiro". Na história de Goethe, o aprendiz de feiticeiro encanta uma vassoura para ir buscar água, mas depressa se apercebe de que não sabe como fazer a vassoura parar. Tenta então cortá-la com um machado, o que apenas resulta em mais vassouras e mais água. A preocupação é que uma máquina inteligente possa agir de forma semelhante: faz aquilo que lhe pedimos, mas quando pedimos que pare, interpreta esse novo pedido como um obstáculo à execução do pedido original. A máquina fará tudo ao seu alcance para prosseguir o objetivo inicial. Uma ilustração frequentemente citada do problema do desalinhamento de objetivos é pedir a uma máquina que maximize a produção de clipes. Uma vez iniciado esse objetivo, nada a conseguiria parar. A máquina converteria todos os recursos da Terra em clipes.

A ameaça do desalinhamento de objetivos assenta em duas improbabilidades: primeiro, embora a máquina aceite o nosso

pedido inicial, ignora todos os pedidos subsequentes; segundo, a máquina é capaz de se apropriar de recursos suficientes para impedir todos os esforços humanos de a travar.

Como tenho salientado repetidamente, a inteligência é a capacidade de aprender um modelo do mundo. Tal como um mapa, esse modelo pode indicar como atingir um determinado resultado, mas, por si só, não possui desejos nem impulsos. Somos nós, os criadores de máquinas inteligentes, que temos de fazer um esforço consciente para lhes incorporar motivações. Por que razão haveríamos de conceber uma máquina que aceitasse o nosso primeiro pedido e ignorasse todos os seguintes? Isso seria tão plausível como projetar um carro autónomo que, depois de lhe dizermos o destino, ignorasse qualquer pedido posterior para parar ou alterar a rota. Para piorar, seria como se esse carro fosse concebido para trancar as portas e desligar o volante, os travões, o botão de ignição, etc. Note-se que um carro autónomo não desenvolverá objetivos por iniciativa própria. Naturalmente, alguém poderia desenhar um carro que perseguisse os seus próprios objetivos e ignorasse os pedidos humanos. Um carro assim poderia causar danos. Mas, mesmo nesse caso, não constituiria uma ameaça existencial sem que se verificasse também o segundo requisito.

O segundo requisito do risco de desalinhamento de objetivos é que uma máquina inteligente consiga apropriar-se dos recursos da Terra para concretizar os seus objetivos, ou, de outras formas, impedir-nos de a travar. É difícil imaginar como tal poderia acontecer. Para isso, a máquina teria de controlar a vasta maioria das comunicações, da produção e dos transportes do mundo. É evidente que um carro inteligente descontrolado não tem essa capacidade. Um cenário possível em que uma máquina inteligente nos impediria de a desligar seria através de chantagem. Por exemplo, se colocássemos uma máquina inteligente no comando de armamento nuclear, ela poderia afirmar: “Se tentarem parar-me, explodirei tudo.” Ou, se uma máquina controlasse a maior parte da internet, poderia ameaçar com todo o tipo de caos, perturbando as comunicações e o comércio.

Temos preocupações semelhantes relativamente aos humanos. É por isso que nenhuma pessoa ou entidade isolada pode controlar toda a internet, e por que razão é necessário o acordo de várias pessoas para lançar um míssil nuclear. Máquinas inteligentes não desenvolverão objetivos desalinhados a menos que nos esforcemos deliberadamente por lhes conferir essa capacidade. E mesmo que o fizéssemos, nenhuma máquina poderia apoderar-se dos recursos mundiais sem que lho permitíssemos. Não permitimos que um único ser humano, ou mesmo um pequeno grupo de pessoas, controle os recursos do planeta. Temos de ser igualmente cautelosos no caso das máquinas.

O Contra-Argumento

Estou convicto de que as máquinas inteligentes não representam uma ameaça existencial para a Humanidade. Um contra-argumento comum, apresentado por quem discorda, é o seguinte: os povos indígenas, ao longo da História, sentiam-se igualmente seguros. Mas, quando estrangeiros surgiram com armas e tecnologias superiores, essas populações foram subjugadas e destruídas. Nós, afirmam, estamos igualmente vulneráveis e não podemos confiar na nossa sensação de segurança. Não conseguimos imaginar quão mais inteligentes, rápidas e capazes as máquinas poderão ser em relação a nós e, por isso, estamos em risco.

Há alguma verdade neste argumento. Algumas máquinas inteligentes serão de facto mais inteligentes, rápidas e capazes do que os humanos. Mas a questão essencial regressa à motivação. As máquinas inteligentes quererão apoderar-se da Terra, submeter-nos ou fazer algo que nos possa prejudicar? A destruição das culturas indígenas resultou das motivações dos invasores — entre elas a ganância, a fama e o desejo de dominação. Estes são impulsos do cérebro primitivo. A tecnologia superior ajudou os invasores, mas não foi a causa principal do massacre.

Mais uma vez, as máquinas inteligentes não possuirão emoções nem impulsos humanos, a menos que lhos incutamos deliberadamente. Desejos, objetivos e agressividade não surgem por magia quando algo se torna inteligente. Para apoiar este ponto

de vista, considere-se que a maior perda de vidas indígenas não foi causada diretamente pelos invasores humanos, mas por doenças introduzidas — bactérias e vírus contra os quais as populações nativas tinham pouca ou nenhuma defesa. Os verdadeiros assassinos foram organismos simples, com o único impulso de se multiplicarem, e sem qualquer tecnologia avançada. A inteligência, nesse caso, teve um álibi: não esteve presente na maioria do genocídio.

Acredito que a autorreplicação constitui uma ameaça muito maior para a Humanidade do que a inteligência das máquinas. Se alguém mal-intencionado quisesse criar algo para eliminar todos os seres humanos, a forma mais eficaz seria conceber novos vírus e bactérias altamente infeciosos e contra os quais o nosso sistema imunitário não pudesse defender-se. Em teoria, seria possível que uma equipa descontrolada de cientistas e engenheiros criasse máquinas inteligentes com o desejo de se autorreplicarem. Essas máquinas teriam também de ser capazes de se copiar sem interferência humana. Mas estes eventos parecem altamente improváveis e, mesmo que ocorressem, nada disso aconteceria de forma rápida. O ponto essencial é este: tudo o que for capaz de se autorreplicar — especialmente vírus e bactérias — constitui uma potencial ameaça existencial. A inteligência, por si só, não.

Não podemos conhecer o futuro e, por isso, não conseguimos antecipar todos os riscos associados à inteligência artificial, tal como não podemos antecipar todos os riscos de qualquer outra nova tecnologia. Mas, à medida que avançamos e debatemos os riscos e benefícios da inteligência artificial, recomendo que se reconheça a distinção entre três coisas: replicação, motivações e inteligência.

- **Replicação:** Tudo o que é capaz de se autorreplicar é perigoso. A Humanidade pode ser dizimada por um vírus biológico. Um vírus informático pode colapsar a internet. Máquinas inteligentes não terão a capacidade nem o desejo de se autorreplicarem, a menos que os humanos se esforcem deliberadamente para que assim seja.

- **Motivações:** As motivações e impulsos biológicos são fruto da evolução. A evolução descobriu que os animais com certos impulsos se replicavam melhor do que outros. Uma máquina que não se replica nem evolui não irá, de forma súbita, desenvolver o desejo de dominar ou escravizar os outros.
- **Inteligência:** Das três, a inteligência é a mais inofensiva. Uma máquina inteligente não começará, por si só, a autorreplicar-se, nem desenvolverá espontaneamente impulsos e motivações. Teremos de nos empenhar conscientemente em projetar as motivações que queremos que as máquinas inteligentes possuam. Mas, a menos que estas se autorrepliquem e evoluam, não representarão por si mesmas um risco existencial para a Humanidade.

Não quero deixar a impressão de que a inteligência artificial não é perigosa. Tal como qualquer tecnologia poderosa, pode causar grandes danos se for usada por humanos com intenções maléficas. Voltemos a imaginar milhões de armas autónomas inteligentes, ou o uso de máquinas inteligentes para fins de propaganda e controlo político. O que deveremos fazer perante isto? Deveríamos impor uma proibição à investigação e desenvolvimento da IA? Seria difícil — e poderia também ir contra os nossos melhores interesses. A inteligência artificial trará enormes benefícios à sociedade e, como defenderei na próxima secção do livro, poderá ser necessária à nossa sobrevivência a longo prazo. Por agora, parece que a melhor opção é trabalhar arduamente para estabelecer acordos internacionais vinculativos sobre o que é aceitável e o que não é — à semelhança do que fazemos com as armas químicas.

A inteligência artificial é frequentemente comparada a um génio dentro de uma garrafa: uma vez libertado, já não pode ser lá colocado de novo, e perderemos rapidamente a capacidade de o controlar. O que tentei mostrar neste capítulo é que esses receios não têm fundamento. Não vamos perder o controlo, e nada acontecerá de forma súbita, como receiam os defensores da explosão de inteligência. Se começarmos agora, teremos tempo

mais do que suficiente para ponderar os riscos e benefícios e decidir como queremos avançar.

Na próxima e última secção do livro, iremos explorar os riscos existenciais — e as oportunidades — da inteligência humana.

PARTE 3

INTELIGÊNCIA HUMANA

Estamos num ponto de inflexão da história da Terra — um período de mudanças rápidas e dramáticas, tanto no planeta como nas formas de vida que o habitam. O clima está a mudar a um ritmo tão acelerado que é provável que, nos próximos cem anos, algumas cidades se tornem inabitáveis e vastas regiões agrícolas se tornem estéreis. As espécies estão a extinguir-se a um ritmo tão elevado que alguns cientistas consideram estar em curso o sexto grande evento de extinção na história da Terra. A inteligência humana é a causa destas mudanças súbitas.

A vida surgiu na Terra há cerca de 3,5 mil milhões de anos. Desde o início, o rumo da vida foi governado pelos genes e pela evolução. Não há plano nem direção desejada na evolução. As espécies evoluíam e extinguiam-se conforme a sua capacidade de deixar descendência com cópias dos seus genes. A vida era movida pela sobrevivência competitiva e pela reprodução. Nada mais importava.

A nossa inteligência permitiu à nossa espécie, o *Homo sapiens*, prosperar e ter sucesso. Em apenas algumas centenas de anos — um instante, em termos geológicos — duplicámos a nossa esperança média de vida, curámos muitas doenças e eliminámos a fome para a vasta maioria da população humana. Vivemos com mais saúde, maior conforto e menos esforço do que os nossos antepassados.

Os seres humanos são inteligentes há centenas de milhares de anos — então, por que razão esta mudança repentina na nossa sorte? O que é novo é o surgimento recente e rápido da nossa tecnologia e das nossas descobertas científicas, que nos permitiram produzir alimentos em abundância, eliminar doenças e transportar bens para onde são mais necessários.

Mas com o nosso sucesso vieram também os problemas. A nossa população passou de mil milhões, há duzentos anos, para

perto de oito mil milhões atualmente. Somos tantos que estamos a poluir todas as partes do planeta. É agora evidente que o nosso impacto ecológico é tão severo que, no mínimo, deslocará centenas de milhões de pessoas; no pior cenário, tornará a Terra inabitável. O clima não é a nossa única preocupação. Algumas das nossas tecnologias, como as armas nucleares e a edição genética, têm o potencial de permitir que um pequeno número de pessoas mate milhares de milhões de outras.

A nossa inteligência tem sido a fonte do nosso sucesso, mas tornou-se também numa ameaça existencial. A forma como agirmos nos próximos anos determinará se esta ascensão repentina levará a um colapso igualmente súbito — ou, alternativamente, se sairemos deste período de mudança acelerada numa trajetória sustentável. Estes são os temas que abordo nos capítulos restantes do livro.

Começo por analisar os riscos inerentes à nossa inteligência e à estrutura do nosso cérebro. A partir desta base, discuto várias opções que poderemos seguir para aumentar as probabilidades de sobrevivência a longo prazo. Examino iniciativas e propostas já existentes, considerando-as à luz da teoria do cérebro. E apresento novas ideias que, na minha opinião, merecem ser consideradas, mas que, tanto quanto sei, ainda não entraram no discurso dominante.

O meu objetivo não é prescrever o que devemos fazer, mas sim incentivar conversas sobre questões que, a meu ver, não estão a ser suficientemente debatidas. A nossa nova compreensão do cérebro oferece-nos a oportunidade de reavaliar os riscos e as oportunidades que enfrentamos. Algumas das ideias que apresento poderão ser algo controversas, mas esse não é o meu propósito. Procuro fazer uma avaliação honesta e imparcial da situação em que nos encontramos — e explorar o que poderemos fazer em relação a ela.

CAPÍTULO 12

Crenças Falsas

Quando éramos adolescentes, os meus amigos e eu estávamos fascinados com a hipótese do cérebro numa cuba. Seria possível que os nossos cérebros estivessem mergulhados numa cuba de nutrientes que os mantinha vivos, enquanto as suas entradas e saídas estivessem ligadas a um computador? A hipótese do cérebro numa cuba sugere a possibilidade de que o mundo em que pensamos viver possa não ser o mundo real, mas sim um mundo falso, simulado por um computador. Embora eu não acredite que os nossos cérebros estejam ligados a um computador, o que está a acontecer é quase tão estranho. O mundo em que pensamos viver não é o mundo real; é uma simulação do mundo real. E isto conduz a um problema: aquilo em que acreditamos é, frequentemente, falso.

O seu cérebro está dentro de uma caixa — o crânio. Não há sensores no cérebro propriamente dito, pelo que os neurónios que o constituem estão sentados na escuridão, isolados do mundo exterior. A única forma de o cérebro saber algo sobre a realidade é através das fibras nervosas sensoriais que entram no crânio. As fibras nervosas provenientes dos olhos, ouvidos e pele têm todas o mesmo aspeto, e os impulsos que viajam por elas são idênticos. Não entra luz nem som no crânio — apenas impulsos elétricos.

O cérebro envia também fibras nervosas aos músculos, os quais movimentam o corpo e os seus sensores, alterando assim a parte do mundo que o cérebro está a captar. Através da repetição contínua de sentir e mover, sentir e mover, o cérebro aprende um modelo do mundo exterior ao crânio.

Repare, mais uma vez, que não entra luz, toque ou som no cérebro. Nenhuma das perceções que constituem a nossa experiência mental — da suavidade de um animal de estimação, ao suspiro de um amigo, às cores das folhas no Outono — chega até nós através dos nervos sensoriais. Os nervos apenas transmitem

impulsos elétricos. E como não temos perceção consciente desses impulsos, tudo aquilo que de facto percebemos tem de ser fabricado no cérebro. Mesmo as sensações mais básicas de luz, som e toque são criações do cérebro — existem apenas no seu modelo do mundo.

Poderá opor-se a esta caracterização. Afinal, os impulsos nervosos de entrada não representam, de algum modo, a luz e o som? De certa forma. Existem propriedades no universo — como a radiação eletromagnética e as ondas de compressão das moléculas do ar — que conseguimos detetar. Os nossos órgãos sensoriais convertem essas propriedades em impulsos nervosos, os quais são então convertidos na nossa perceção de luz e som. Mas os órgãos sensoriais não captam tudo. Por exemplo, a luz, no mundo real, abrange um vasto espectro de frequências, mas os nossos olhos só são sensíveis a uma pequena fração desse espectro. Do mesmo modo, os nossos ouvidos só detetam sons dentro de uma gama estreita de frequências. Assim, a nossa perceção de luz e som só pode representar uma parte daquilo que está a acontecer no universo. Se pudéssemos captar todas as frequências da radiação eletromagnética, veríamos transmissões de rádio, radares, e teríamos visão por raios-X. Com sensores diferentes, o mesmo universo daria origem a experiências percetivas diferentes.

Há dois pontos essenciais a reter: o cérebro apenas tem acesso a uma parte limitada do mundo real, e aquilo que percebemos é o nosso modelo do mundo — não o mundo em si. Neste capítulo, exploro de que forma estas ideias conduzem a crenças falsas — e o que, se é que algo, podemos fazer em relação a isso.

Vivemos numa Simulação

Num dado momento, alguns neurónios no cérebro estão ativos e outros não. Os neurónios ativos representam aquilo que estamos a pensar e a perceber naquele instante. Importa sublinhar que estes pensamentos e perceções são relativos ao modelo que o cérebro tem do mundo — não ao mundo físico fora do crânio. Assim, o mundo que percebemos é uma simulação do mundo real.

Eu sei que não parece que estamos a viver numa simulação. Parece que estamos a olhar diretamente para o mundo, a tocá-lo, a cheirá-lo e a senti-lo. Por exemplo, é comum pensar-se que os olhos funcionam como uma câmara — que o cérebro recebe uma imagem dos olhos e que essa imagem é o que vemos. Embora esta forma de pensar seja natural, não é verdadeira. Recorde que, anteriormente no livro, expliquei como a nossa perceção visual é estável e uniforme, mesmo quando os sinais recebidos dos olhos são distorcidos e estão em constante mudança. A verdade é que o que percebemos é o nosso modelo do mundo, não o mundo em si, nem os impulsos que entram no crânio em fluxo contínuo. No decurso do nosso dia, os estímulos sensoriais ativam as partes adequadas do modelo do mundo no cérebro — mas aquilo que percebemos e acreditamos estar a acontecer é esse modelo. A nossa realidade é semelhante à hipótese do cérebro numa cuba: vivemos num mundo simulado, mas não por um computador — sim dentro da nossa cabeça.

Esta ideia é tão contraintuitiva que vale a pena explorá-la com alguns exemplos. Comecemos com a perceção de localização. Uma fibra nervosa que transmite a sensação de pressão na ponta de um dedo não transporta qualquer informação sobre a posição do dedo. Essa fibra responde exatamente da mesma maneira esteja o dedo a tocar em algo à sua frente ou ao seu lado. No entanto, você percebe essa sensação de toque como estando num determinado local em relação ao seu corpo. Isto parece tão natural que provavelmente nunca se perguntou como tal acontece. Como referi anteriormente, a explicação está nas colunas corticais que representam cada parte do corpo. E dentro dessas colunas, há neurónios que codificam a localização de cada parte corporal. Percebe o seu dedo como estando num determinado local porque as células que representam essa localização o indicam.

O modelo pode estar errado. Por exemplo, pessoas que perderam um membro frequentemente continuam a sentir que esse membro ainda existe. O modelo do cérebro ainda inclui o membro perdido e a sua localização. Assim, mesmo que fisicamente já não esteja presente, o indivíduo continua a percebê-lo como se estivesse ali, ligado ao corpo. O "membro fantasma" pode inclusive "mover-se" para diferentes posições. Amputados podem relatar que

o seu braço desaparecido está encostado ao corpo, ou que a sua perna desaparecida está dobrada ou esticada. Podem sentir comichão ou dor em locais específicos do membro ausente. Estas sensações são "lá fora", no local onde o membro é percebido — mas, fisicamente, nada existe ali. O modelo do cérebro inclui o membro e, esteja certo ou errado, é isso que é percebido.

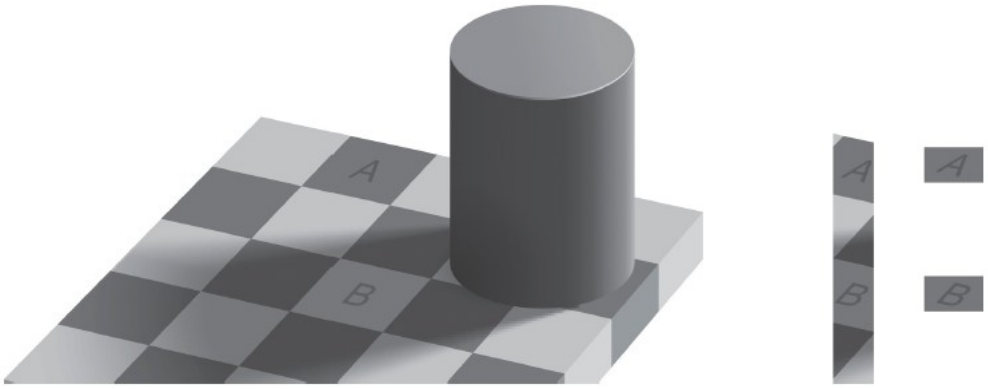
Algumas pessoas apresentam o problema oposto. Têm um membro perfeitamente funcional, mas sentem que não lhes pertence. Por parecer estranho ou "alheio", podem mesmo desejar a sua amputação. A razão pela qual certas pessoas sentem que uma parte do corpo não lhes pertence ainda é desconhecida, mas esta perceção falsa está claramente enraizada no facto de o modelo cerebral não representar o membro de forma normal. Se o modelo do seu cérebro não incluir uma perna esquerda, então essa perna será percebida como algo estranho ao seu corpo. Seria como se alguém colasse uma chávena ao seu cotovelo — quereria livrar-se dela o mais depressa possível.

Mesmo uma pessoa com um corpo absolutamente normal pode ser enganada na perceção que tem dele. A ilusão da mão de borracha é um truque clássico, em que o sujeito vê uma mão de borracha, mas não a sua verdadeira mão. Quando outra pessoa acaricia de forma idêntica a mão de borracha visível e a mão real escondida, o sujeito começa a sentir que a mão de borracha é parte do seu corpo.

Estes exemplos mostram-nos que o modelo que o cérebro constrói do mundo pode estar errado. Podemos perceber coisas que não existem (como o membro fantasma), e podemos perceber de forma errada coisas que existem (como o membro "alheio" ou a mão de borracha). Estes são casos em que o modelo do cérebro está claramente errado — e de forma prejudicial. Por exemplo, a dor fantasma pode ser incapacitante. No entanto, é bastante comum que o modelo cerebral não coincida com os sinais sensoriais recebidos. E, na maioria dos casos, isso é útil.

A imagem seguinte, criada por Edward Adelson, é um exemplo poderoso da diferença entre o modelo cerebral do mundo (aquilo que percebe) e aquilo que realmente é captado pelos sentidos. Na figura à esquerda, o quadrado com a letra A parece mais escuro do

que o quadrado com a letra B. Contudo, os quadrados A e B são absolutamente idênticos. Pode estar a dizer para si próprio: "Isso não é possível. O A é claramente mais escuro do que o B." Mas estará enganado. A melhor forma de confirmar que A e B têm exatamente o mesmo tom é ocultar todas as outras partes da imagem e deixar apenas esses dois quadrados visíveis — verá então que são idênticos. Para o ajudar, incluo dois recortes da imagem principal. O efeito é menos evidente na faixa recortada e desaparece completamente quando vê apenas os quadrados A e B.



Para chamar a isto uma ilusão seria sugerir que o cérebro está a ser enganado, mas acontece precisamente o contrário. O seu cérebro está a perceber corretamente um tabuleiro de xadrez e não está a ser iludido pela sombra. Um padrão de tabuleiro de xadrez é um padrão de tabuleiro de xadrez, independentemente de ter ou não uma sombra sobre ele. O modelo do cérebro diz que os tabuleiros de xadrez têm quadrados alternadamente escuros e claros, e é isso que percebe, mesmo que, neste caso, a luz proveniente de um quadrado "escuro" e de um quadrado "claro" seja idêntica.

O modelo do mundo que reside no nosso cérebro é geralmente exato. Costuma captar a estrutura da realidade de forma independente da nossa perspetiva momentânea ou de outros dados contraditórios, como a sombra sobre o tabuleiro de xadrez. No entanto, o modelo do mundo do cérebro pode também estar redondamente enganado.

Crenças Falsas

Uma crença falsa é quando o modelo do cérebro acredita que algo existe, quando na realidade esse algo não existe no mundo físico. Pense novamente nos membros fantasmas. Um membro fantasma ocorre porque existem colunas no neocórtex que modelam esse membro. Essas colunas contêm neurónios que representam a localização do membro em relação ao corpo. Imediatamente após a remoção do membro, essas colunas continuam lá, e continuam a conter um modelo do membro. Por isso, a pessoa acredita que o membro ainda se encontra numa determinada posição, mesmo que ele já não exista no mundo físico. O membro fantasma é um exemplo de crença falsa. (A perceção do membro fantasma tende a desaparecer ao fim de alguns meses, à medida que o cérebro ajusta o seu modelo do corpo, mas em algumas pessoas pode persistir durante anos.)

Consideremos agora outro modelo falso. Algumas pessoas acreditam que a Terra é plana. Durante dezenas de milhares de anos, todas as experiências humanas foram consistentes com a ideia de que a Terra é plana. A curvatura da Terra é tão subtil que, ao longo de uma vida, não era possível detetá-la. Existem algumas inconsistências subtis, como o facto de o casco de um navio desaparecer no horizonte antes dos mastros, mas isso é difícil de observar mesmo com uma excelente acuidade visual. Um modelo que diz que a Terra é plana não só é consistente com as nossas sensações, como também é um bom modelo para agir no mundo. Por exemplo, hoje preciso de ir do meu gabinete à biblioteca para devolver um livro. Planear esse trajeto com base num modelo de Terra plana funciona perfeitamente; não preciso de considerar a curvatura da Terra para me deslocar pela cidade. Em termos de sobrevivência quotidiana, um modelo de Terra plana é perfeitamente eficaz — ou, pelo menos, era até há pouco tempo. Hoje, se for astronauta, ou piloto de navios, ou mesmo um viajante internacional frequente, acreditar que a Terra é plana pode ter consequências graves e fatais. Se não for um viajante de longas distâncias, então um modelo de Terra plana continua a funcionar bem no dia-a-dia.

Por que é que algumas pessoas ainda acreditam que a Terra é plana? Como conseguem manter esse modelo em presença de dados sensoriais contrários, como fotografias da Terra tiradas do espaço ou relatos de exploradores que atravessaram o Pólo Sul?

Recorde que o neocórtex está constantemente a fazer previsões. As previsões são a forma como o cérebro testa se o seu modelo do mundo está correto; uma previsão incorreta indica que há algo de errado com o modelo e que ele precisa de ser corrigido. Um erro de previsão gera uma explosão de atividade no neocórtex, que dirige a nossa atenção para o estímulo que causou o erro. Ao focar a atenção nesse estímulo imprevisto, o neocórtex reaprende essa parte do modelo. Isto conduz, por fim, à modificação do modelo cerebral para refletir de forma mais fiel a realidade. A correção de modelos está embutida no neocórtex e, normalmente, funciona de forma fiável.

Para manter um modelo falso, como o da Terra plana, é necessário rejeitar as evidências que contradizem esse modelo. Os crentes da Terra plana dizem que não confiam em nenhuma evidência que não possam experimentar diretamente. Uma fotografia pode ser falsificada. Um relato de um explorador pode ter sido inventado. O envio de pessoas à Lua nos anos 60 pode ter sido uma produção de Hollywood. Se restringir aquilo em que acredita apenas àquilo que pode experienciar diretamente, e se não for astronauta, então um modelo de Terra plana é aquilo com que acabará por ficar. Para manter um modelo falso, também ajuda rodear-se de outras pessoas que partilham as mesmas crenças, tornando mais provável que os estímulos recebidos sejam consistentes com esse modelo. Historicamente, isto implicava isolar-se fisicamente numa comunidade de pessoas com crenças semelhantes, mas hoje em dia é possível alcançar o mesmo resultado selecionando os vídeos que se vê na internet.

Considera agora as alterações climáticas. Existe uma quantidade esmagadora de provas de que a atividade humana está a provocar alterações de grande escala no clima da Terra. Essas alterações, se não forem travadas, poderão conduzir à morte e/ou deslocação de milhares de milhões de pessoas. Há debates legítimos sobre o que devemos fazer em relação às alterações

climáticas, mas há também muitas pessoas que simplesmente negam que estejam a acontecer. O modelo do mundo que possuem diz-lhes que o clima não está a mudar — ou, mesmo que esteja a mudar, não há razão para preocupação.

Como é que os negacionistas das alterações climáticas conseguem manter essa crença falsa perante evidência física substancial? São como os crentes da Terra plana: não confiam na maioria das outras pessoas e baseiam-se apenas naquilo que observam pessoalmente ou naquilo que outras pessoas com a mesma mentalidade lhes dizem. Se não conseguirem ver as alterações climáticas, então não acreditam que estejam a acontecer. Há indícios de que os negacionistas climáticos tendem a mudar de opinião se experienciarem pessoalmente um evento climático extremo ou inundações causadas pela subida do nível do mar.

Se basear a sua visão do mundo apenas nas suas experiências pessoais, então é possível viver uma vida aparentemente normal e acreditar que a Terra é plana, que as alunagens foram forjadas, que a atividade humana não está a alterar o clima global, que as espécies não evoluem, que as vacinas provocam doenças e que os tiroteios em massa são encenados.

Modelos Virais do Mundo

Alguns modelos do mundo são virais, no sentido em que o modelo faz com que o cérebro que o hospeda atue de forma a espalhar esse modelo para outros cérebros. Um modelo de um membro fantasma não é viral; é um modelo incorreto, mas está isolado num único cérebro. Um modelo da Terra plana também não é viral, pois mantê-lo requer confiar apenas nas experiências pessoais. Acreditar que a Terra é plana não leva a agir de modo a espalhar essa crença a outras pessoas.

Modelos virais do mundo prescrevem comportamentos que propagam o modelo de cérebro em cérebro, em número crescente. Por exemplo, o meu modelo do mundo inclui a crença de que todas as crianças devem ter uma boa educação. Se parte dessa educação consistir em ensinar que todas as crianças merecem uma boa

educação, isso inevitavelmente levará a que mais e mais pessoas acreditem que todas as crianças merecem uma boa educação. O meu modelo do mundo — pelo menos a parte sobre a educação universal da infância — é viral. Propagar-se-á a um número cada vez maior de pessoas ao longo do tempo. Mas será ele correto? Isso é difícil de dizer. O meu modelo sobre como os humanos devem comportar-se não é algo físico, como a existência de um membro ou a curvatura da Terra. Outras pessoas possuem um modelo que afirma que só algumas crianças merecem uma boa educação. O modelo delas inclui educar os seus filhos para acreditarem que apenas eles, e pessoas semelhantes a eles, merecem uma boa educação. Este modelo de educação seletiva também é viral — e, argumentavelmente, um modelo mais eficaz na propagação dos genes. Por exemplo, as pessoas que recebem uma boa educação terão melhor acesso a recursos financeiros e a cuidados de saúde, sendo, portanto, mais propensas a transmitir os seus genes do que aquelas com pouca ou nenhuma educação. Do ponto de vista darwiniano, a educação seletiva é uma boa estratégia — desde que os que não a recebem não se revoltem.

Modelos do Mundo Falsos e Virais

Agora voltemo-nos para o tipo mais problemático de modelo do mundo: aqueles que são tanto virais como demonstravelmente falsos. Por exemplo, imaginemos que temos um livro de história com numerosos erros factuais. O livro começa com um conjunto de instruções para o leitor. A primeira instrução diz:

> **"Tudo o que está neste livro é verdadeiro. Ignora qualquer evidência que contradiga este livro."**

A segunda instrução diz:

> **"Se encontrares outras pessoas que também acreditam que este livro é verdadeiro, então deves ajudá-las em tudo o que precisarem, e elas farão o mesmo por ti."**

A terceira instrução diz:

> **"Diz a toda a gente que puderes que este livro é verdadeiro. Se recusarem acreditar, então deves bani-las ou matá-las."**

À partida, poderia pensar: "Quem é que vai acreditar nisto?" Contudo, se apenas os cérebros de algumas pessoas acreditarem que o livro é verdadeiro, então modelos mentais que incluam a veracidade do livro podem espalhar-se de forma viral para muitos outros cérebros ao longo do tempo. O livro não apenas descreve um conjunto de crenças falsas sobre a história, como também prescreve ações específicas. Essas ações levam as pessoas a propagar a crença no livro, a ajudar outros que também acreditam nele e a eliminar fontes de evidência contrária.

Este livro de história é um exemplo de um meme. Introduzido pela primeira vez pelo biólogo Richard Dawkins, um meme é algo que se replica e evolui, de forma semelhante a um gene, mas através da cultura. (Recentemente, o termo meme foi apropriado para designar imagens na internet. Aqui estou a usar a palavra no seu sentido original.) O livro de história é, na verdade, um conjunto de memes que se apoiam mutuamente, do mesmo modo que um organismo individual é constituído por um conjunto de genes interdependentes. Por exemplo, cada instrução individual no livro pode ser considerada um meme.

Os memes do livro de história mantêm uma relação simbiótica com os genes de uma pessoa que acredita nesse livro. Por exemplo, o livro dita que as pessoas que nele acreditam devem receber apoio preferencial por parte de outros crentes. Isto aumenta a probabilidade de os crentes terem mais filhos sobreviventes (mais cópias dos genes), o que, por sua vez, leva a mais pessoas que acreditam que o livro é verdadeiro (mais cópias dos memes).

Memes e genes evoluem, e podem fazê-lo de forma mutuamente reforçadora. Por exemplo, imaginemos que é publicada uma variante do livro de história. A diferença entre a versão original e a nova é a adição de mais algumas instruções no início do livro, como por exemplo:

> **"As mulheres devem ter o maior número de filhos possível" e**
>
> **"Não permitir que as crianças frequentem escolas onde possam ser expostas a críticas ao livro."**

Agora existem dois livros de história em circulação. O livro mais recente, com as suas instruções adicionais, é ligeiramente

mais eficaz na replicação do que o anterior. Assim, com o tempo, dominará. Os genes biológicos dos crentes poderão também evoluir de modo a selecionar pessoas mais dispostas a ter muitos filhos, mais aptas a ignorar evidências que contradigam o livro, ou mais predispostas a prejudicar os não crentes.

Modelos falsos do mundo podem espalhar-se e prosperar enquanto as crenças falsas ajudarem os crentes a propagar os seus genes. O livro de história e os seus crentes estão numa relação simbiótica. Ajudam-se mutuamente a replicar, e evoluem ao longo do tempo de forma mutuamente reforçadora. O livro de história pode estar factualmente errado, mas a vida não se trata de ter um modelo correto do mundo. A vida trata-se de replicação.

Linguagem e a Propagação de Crenças Falsas

Antes da linguagem, o modelo do mundo de um indivíduo estava limitado aos locais por onde ele próprio tinha viajado e às coisas que tinha encontrado pessoalmente. Ninguém podia saber o que havia para lá de uma colina ou do outro lado de um oceano sem lá ir. Aprender sobre o mundo através da experiência pessoal é, em geral, fiável.

Com o advento da linguagem, os seres humanos expandiram o seu modelo do mundo para incluir coisas que não observaram diretamente. Por exemplo, embora eu nunca tenha estado em Havana, posso falar com pessoas que afirmam lá ter estado e ler o que outros escreveram sobre a cidade. Acredito que Havana é um lugar real, porque pessoas em quem confio me dizem que lá estiveram e os seus relatos são consistentes. Hoje, grande parte do que acreditamos sobre o mundo não é diretamente observável, e por isso confiamos na linguagem para aprender sobre esses fenómenos. Isto inclui descobertas como os átomos, as moléculas e as galáxias. Inclui processos lentos como a evolução das espécies e a tectónica de placas. Inclui lugares que nunca visitámos pessoalmente, mas cuja existência aceitamos, como o planeta Neptuno ou, no meu caso, Havana. O triunfo do intelecto humano, o esclarecimento da nossa espécie, reside na expansão do nosso modelo do mundo para além do que conseguimos observar diretamente. Esta expansão do conhecimento tornou-se possível

através de ferramentas — como navios, microscópios e telescópios — e de várias formas de comunicação, como a linguagem escrita e as imagens.

Mas aprender sobre o mundo de forma indireta, através da linguagem, não é totalmente fiável. Por exemplo, é possível que Havana não seja um lugar real. É possível que as pessoas que me falaram de Havana estejam a mentir e a coordenar a sua desinformação para me enganar. O exemplo do falso livro de história mostra como as crenças falsas podem propagar-se através da linguagem mesmo que ninguém esteja intencionalmente a espalhar desinformação.

Existe apenas uma forma conhecida de discernir falsidades de verdades, uma forma de verificar se o nosso modelo do mundo contém erros. Esse método consiste em procurar ativamente evidências que contradigam as nossas crenças. Encontrar evidência que as apoie é útil, mas não conclusivo. Já encontrar evidência contrária é prova de que o modelo na nossa mente está errado e precisa de ser modificado. Procurar ativamente evidências que refutem as nossas crenças é o método científico. É a única abordagem conhecida que nos pode aproximar da verdade.

Hoje, no início do século XXI, crenças falsas proliferam na mente de milhares de milhões de pessoas. Isto é compreensível no caso de mistérios que ainda não foram resolvidos. Por exemplo, é compreensível que há quinhentos anos se acreditasse que a Terra era plana, porque a natureza esférica do planeta não era amplamente compreendida e havia pouca ou nenhuma evidência de que a Terra não era plana. De forma semelhante, é compreensível que hoje existam diferentes crenças sobre a natureza do tempo (e quase todas estarão erradas), visto que ainda não descobrimos o que é o tempo. Mas o que me perturba é o facto de que milhares de milhões de pessoas ainda mantêm crenças que já foram refutadas. Por exemplo, trezentos anos após o início do Iluminismo, a maioria da humanidade continua a acreditar em origens míticas da Terra. Esses mitos de origem foram refutados por montanhas de evidência contrária, e ainda assim, persistem.

Temos as crenças falsas virais como principal culpado disto. Tal como o falso livro de história, os memes dependem de cérebros

para se replicarem e, por isso, evoluíram formas de controlar o comportamento dos cérebros de modo a servir os seus próprios interesses. Como o neocórtex está constantemente a fazer previsões para testar o seu modelo do mundo, esse modelo é, por natureza, autocorretivo. Por si só, um cérebro tenderá inexoravelmente para modelos do mundo cada vez mais precisos. Mas esse processo é travado, à escala global, pelas crenças falsas virais.

Mais adiante, no final do livro, apresentarei uma visão mais otimista da humanidade. Mas antes de nos voltarmos para essa visão mais luminosa, quero falar sobre a ameaça existencial muito real que nós, humanos, representamos para nós próprios.

CAPÍTULO 13

Os Riscos Existenciais da Inteligência Humana

A inteligência, por si só, é benigna. Tal como defendi dois capítulos atrás, a menos que, de forma deliberada, integremos impulsos egoístas, motivações e emoções, as máquinas inteligentes não representarão um risco para a nossa sobrevivência. A inteligência humana, contudo, não é tão benigna. A possibilidade de o comportamento humano conduzir à nossa própria extinção é algo há muito reconhecido. Por exemplo, desde 1947 que o Bulletin of the Atomic Scientists mantém o Relógio do Juízo Final (Doomsday Clock) para alertar sobre quão perto estamos de tornar a Terra inabitável. Inicialmente inspirado pela possibilidade de uma guerra nuclear e da subsequente conflagração poderem destruir o planeta, o Relógio do Juízo Final foi alargado, em 2007, para incluir as alterações climáticas como uma segunda causa potencial de extinção autoinfligida. A questão de saber se as armas nucleares e as alterações climáticas induzidas pelo ser humano constituem ameaças existenciais é debatida, mas não há dúvida de que ambas têm potencial para causar imenso sofrimento humano. No que respeita às alterações climáticas, já ultrapassámos a fase da incerteza; o debate centra-se agora, sobretudo, na gravidade dos seus efeitos, nas populações que serão afetadas, na rapidez do seu avanço e nas medidas que devemos tomar.

As ameaças existenciais representadas pelas armas nucleares e pelas alterações climáticas não existiam há cem anos. Dada a velocidade atual da mudança tecnológica, é quase certo que criaremos novas ameaças existenciais nos próximos anos. Precisamos de combater essas ameaças, mas, se queremos ter sucesso a longo prazo, temos de olhar para estes problemas sob uma perspetiva sistémica. Neste capítulo, concentro-me nos dois riscos sistémicos fundamentais associados ao cérebro humano.

O primeiro está ligado às partes mais antigas do nosso cérebro. Embora o neocórtex nos confira uma inteligência superior, cerca de 30 por cento do nosso cérebro evoluiu muito antes, e dá origem aos nossos desejos e comportamentos mais primitivos. O nosso neocórtex inventou tecnologias poderosas, capazes de transformar a Terra inteira, mas o comportamento humano que dirige estas tecnologias transformadoras é frequentemente dominado pelo cérebro antigo — egoísta e míope.

O segundo risco está mais diretamente associado ao neocórtex e à inteligência. O neocórtex pode ser enganado. Pode formar crenças erróneas sobre aspetos fundamentais da realidade. Com base nessas crenças falsas, podemos agir contra os nossos próprios interesses a longo prazo.

Os Riscos do Cérebro Antigo

Somos animais, descendentes de inúmeras gerações de outros animais. Cada um dos nossos antepassados teve êxito em gerar, pelo menos, um descendente — que, por sua vez, também teve pelo menos um descendente — e assim sucessivamente. A nossa linhagem remonta a milhares de milhões de anos. Ao longo de todo este vasto percurso temporal, a medida última do sucesso — provavelmente a única — foi a transmissão preferencial dos próprios genes à geração seguinte.

Os cérebros eram úteis apenas na medida em que aumentavam a sobrevivência e a fecundidade do animal que os possuía. Os primeiros sistemas nervosos eram simples; limitavam-se a controlar reações reflexas e funções corporais. A sua estrutura e funcionamento eram inteiramente determinados pelos genes. Com o tempo, as funções incorporadas expandiram-se para incluir comportamentos que hoje consideramos desejáveis, como cuidar da prole e cooperar socialmente. Mas também surgiram comportamentos que vemos com menor benevolência, como disputar território, lutar por direitos de acasalamento, cópula forçada e roubo de recursos.

Todos os comportamentos inatos, independentemente de os considerarmos desejáveis ou não, surgiram porque foram

adaptações bem-sucedidas. As partes mais antigas do nosso cérebro ainda albergam esses comportamentos primitivos; todos vivemos com esse legado. Naturalmente, cada um de nós situa-se num ponto do espectro entre o quanto expressamos esses comportamentos do cérebro antigo e o quanto o nosso neocórtex, mais lógico, é capaz de os controlar. Acredita-se que parte dessa variação seja genética. Quanto do resto será cultural, é algo ainda desconhecido.

Portanto, apesar de sermos inteligentes, o nosso cérebro antigo continua presente. E continua a operar segundo as regras estabelecidas por centenas de milhões de anos de sobrevivência. Continuamos a lutar por território, a disputar direitos de acasalamento, e continuamos a enganar, a violar e a trair os nossos semelhantes. Nem todos fazem estas coisas, e ensinamos às nossas crianças os comportamentos que desejamos que adotem, mas basta um olhar rápido sobre as notícias de qualquer dia para confirmar que, enquanto espécie, atravessando culturas e comunidades, ainda não conseguimos libertar-nos destes comportamentos primitivos menos desejáveis. Mais uma vez, quando me refiro a um comportamento como sendo menos desejável, falo do ponto de vista individual ou societal. Do ponto de vista dos genes, todos esses comportamentos são úteis.

Por si só, o cérebro antigo não representa um risco existencial. Os comportamentos do cérebro antigo foram, afinal, adaptações bem-sucedidas. No passado, se, na luta por território, uma tribo eliminava todos os membros de outra tribo, isso não punha em risco toda a humanidade. Havia vencedores e vencidos. As ações de uma ou algumas pessoas estavam confinadas a uma parte do globo e a uma parte da humanidade. Hoje, o cérebro antigo representa uma ameaça existencial porque o nosso neocórtex criou tecnologias capazes de transformar — e até destruir — todo o planeta. As ações míopes do cérebro antigo, aliadas às tecnologias de alcance global criadas pelo neocórtex, tornaram-se uma ameaça existencial para a humanidade. Vejamos como isto se manifesta atualmente, analisando as alterações climáticas e uma das suas causas subjacentes: o crescimento populacional.

Crescimento Populacional e Alterações Climáticas

As alterações climáticas causadas pelo ser humano resultam de dois fatores. Um é o número de pessoas que vivem na Terra; o outro é a quantidade de poluição que cada pessoa gera. Ambos os números estão a aumentar. Vejamos o crescimento populacional.

Em 1960, havia cerca de três mil milhões de pessoas no planeta. As minhas primeiras memórias datam dessa década. Não me recordo de alguém ter sugerido, nessa altura, que os problemas do mundo se resolveriam se houvesse o dobro da população. Hoje, a população humana aproxima-se dos oito mil milhões e continua a crescer.

A lógica simples indica que a Terra estaria menos sujeita a formas de degradação e colapso causadas pelo ser humano se houvesse menos pessoas. Por exemplo, se existissem dois mil milhões de pessoas em vez de oito mil milhões, é possível que os ecossistemas terrestres conseguissem absorver o nosso impacto sem mudanças rápidas e radicais. Mesmo que a Terra não pudesse sustentar, de forma duradoura, dois mil milhões de seres humanos, teríamos mais tempo para ajustar os nossos comportamentos e viver de maneira sustentável.

Por que é que, então, a população da Terra passou de três mil milhões em 1960 para os atuais oito mil milhões? Por que não se manteve nos três mil milhões, ou até diminuiu para dois mil milhões? Quase toda a gente concordaria que o planeta estaria em melhor situação com menos pessoas, não com mais. Por que não é isso que está a acontecer? A resposta pode parecer óbvia, mas vale a pena analisá-la mais atentamente.

A vida baseia-se numa ideia muito simples: os genes fazem o maior número possível de cópias de si próprios. Isso levou os animais a tentarem ter o máximo de descendência possível, e as espécies a tentarem ocupar o maior número de habitats possível. Os cérebros evoluíram para servir este aspeto mais básico da vida. Os cérebros ajudam os genes a produzirem mais cópias de si mesmos.

Contudo, o que é benéfico para os genes nem sempre o é para os indivíduos. Por exemplo, do ponto de vista de um gene, não há problema se uma família humana tiver mais filhos do que aqueles que consegue alimentar. É certo que, em alguns anos, as crianças poderão morrer à fome, mas noutros anos isso não acontecerá. Do ponto de vista do gene, é preferível, ocasionalmente, ter filhos a mais do que filhos a menos. Algumas crianças sofrerão horrivelmente, e os pais enfrentarão dificuldades e luto, mas os genes não se importam. Nós, enquanto indivíduos, existimos para servir os interesses dos genes. Os genes que nos levam a ter o maior número possível de filhos serão os mais bem-sucedidos, mesmo que isso, por vezes, conduza à morte e ao sofrimento.

De forma semelhante, do ponto de vista do gene, é vantajoso que os animais tentem viver em novos locais, mesmo que muitas dessas tentativas fracassem. Imaginemos que uma tribo humana se divide e ocupa quatro novos habitats, mas apenas um desses subgrupos sobrevive, enquanto os outros três lutam, passam fome e acabam por desaparecer. Haverá muito sofrimento para os indivíduos humanos, mas será um sucesso para o gene, que agora ocupa o dobro do território que ocupava antes.

Os genes não compreendem nada. Não desfrutam de ser genes, nem sofrem quando falham em replicar-se. São simplesmente moléculas complexas, capazes de se replicar.

O neocórtex, por outro lado, compreende o quadro geral. Ao contrário do cérebro antigo — com os seus objetivos e comportamentos programados —, o neocórtex aprende um modelo do mundo e é capaz de prever as consequências do crescimento populacional descontrolado. Assim, conseguimos antecipar a miséria e o sofrimento que enfrentaremos se continuarmos a permitir que a população da Terra cresça. Por que não estamos, então, a reduzir coletivamente a população? Porque o cérebro antigo continua a comandar.

Recordemos o exemplo de um pedaço de bolo tentador, mencionado no Capítulo 2. O nosso neocórtex pode saber que comer bolo nos faz mal, que pode levar à obesidade, a doenças e à morte prematura. Podemos sair de casa pela manhã determinados a comer apenas alimentos saudáveis. Contudo, ao ver e cheirar o

bolo, muitas vezes comemo-lo na mesma. O cérebro antigo está no comando, e esse cérebro evoluiu num tempo em que as calorias eram escassas. O cérebro antigo não conhece as consequências futuras. Na batalha entre o cérebro antigo e o neocórtex, o cérebro antigo costuma vencer. Comemos o bolo.

Dado que temos dificuldade em controlar o que comemos, fazemos o que está ao nosso alcance. Usamos a inteligência para mitigar os danos. Criamos intervenções médicas, como medicamentos e cirurgias. Organizamos conferências sobre a epidemia de obesidade. Desenvolvemos campanhas para educar as pessoas sobre os riscos de uma alimentação pouco saudável. Mas, embora logicamente fosse melhor simplesmente comermos melhor, o problema fundamental mantém-se. Continuamos a comer o bolo.

Algo semelhante está a acontecer com o crescimento populacional. Sabemos que, a certa altura, teremos de parar o crescimento da população. Isto é pura lógica; as populações não podem crescer indefinidamente, e muitos ecologistas acreditam que a nossa já é insustentável. Mas temos dificuldade em controlar a população porque o cérebro antigo quer ter filhos. Assim, em vez disso, usamos a nossa inteligência para melhorar dramaticamente a agricultura, inventando novas culturas e novos métodos para aumentar a produtividade. Criámos também tecnologias que permitem enviar alimentos para qualquer parte do mundo. Usando a nossa inteligência, alcançámos o milagre: reduzimos a fome e a escassez alimentar numa época em que a população humana quase triplicou. Contudo, isso só pode durar até certo ponto. Ou o crescimento populacional pára, ou, em algum momento no futuro, haverá grande sofrimento humano na Terra. Isso é uma certeza.

É claro que esta situação não é tão a preto e branco como aqui a apresentei. Algumas pessoas decidem, com lógica, ter menos filhos ou nenhum; outras podem não ter a educação necessária para compreender as ameaças a longo prazo das suas ações; e muitas são tão pobres que dependem de ter filhos para sobreviver. As questões associadas ao crescimento populacional são complexas, mas se recuarmos e olharmos para o quadro geral, veremos que os seres humanos compreendem a ameaça do crescimento populacional há, pelo menos, cinquenta anos — e, nesse tempo, a

nossa população quase triplicou. Na raiz deste crescimento estão as estruturas do cérebro antigo e os genes que elas servem. Felizmente, existem formas de o neocórtex vencer esta batalha.

Como o Neocórtex Pode Frustrar o Cérebro Antigo

O curioso no que respeita à sobrepopulação é que a ideia de haver uma população humana mais reduzida não é, em si, controversa — mas falar sobre como poderíamos alcançá-la a partir da situação atual é social e politicamente inaceitável. Talvez nos recordemos da amplamente criticada política do filho único na China. Talvez associemos inconscientemente a redução populacional a genocídio, eugenia ou pogroms. Por qualquer razão que seja, a ideia de deliberadamente procurar uma população menor raramente é discutida. Com efeito, quando a população de um país está em declínio, como acontece hoje no Japão, isso é considerado uma crise económica. É raro ouvir a população decrescente do Japão ser apresentada como um modelo para o resto do mundo.

Temos a sorte de haver uma solução simples e engenhosa para o crescimento populacional — uma solução que não obriga ninguém a fazer o que não quer, uma solução que sabemos ser eficaz na redução da população para um nível mais sustentável, e uma solução que, além disso, aumenta a felicidade e o bem-estar das pessoas envolvidas. No entanto, é uma solução à qual muitas pessoas ainda se opõem. Essa solução simples e engenhosa consiste em garantir que todas as mulheres tenham a capacidade de controlar a sua própria fertilidade — e sejam capacitadas para exercer essa escolha, se assim o desejarem.

Chamo-lhe uma solução engenhosa porque, na batalha entre o cérebro antigo e o neocórtex, o cérebro antigo quase sempre vence. A invenção dos métodos contracetivos mostra como o neocórtex pode usar a sua inteligência para assumir o controlo.

Os genes propagam-se melhor quando temos o maior número possível de descendentes. O desejo sexual é o mecanismo que a evolução desenvolveu para servir os interesses dos genes. Mesmo

que não queiramos ter mais filhos, é difícil deixar de ter relações sexuais. Por isso, utilizámos a nossa inteligência para criar métodos de contraceção que permitem ao cérebro antigo ter tanto sexo quanto deseje, sem gerar mais filhos. O cérebro antigo não é inteligente; não compreende o que está a fazer nem porquê. O nosso neocórtex, com o seu modelo do mundo, consegue ver os inconvenientes de ter filhos a mais e reconhecer as vantagens de adiar a constituição de família. Em vez de lutar contra o cérebro antigo, o neocórtex permite-lhe obter aquilo que deseja, evitando o resultado final indesejado.

Por que é que, então, existe ainda tanta resistência à capacitação das mulheres? Por que há tantas pessoas que se opõem à igualdade salarial, ao acesso universal a creches e ao planeamento familiar? E por que é que as mulheres continuam a encontrar obstáculos ao alcançar uma representação equitativa em posições de poder? Por quase todos os critérios objetivos, capacitar as mulheres conduzirá a um mundo mais sustentável e com menos sofrimento humano. Visto de fora, parece contraproducente opor-se a isso. Podemos atribuir este dilema ao cérebro antigo — e a crenças falsas que se disseminam como vírus. E isto leva-nos ao segundo risco fundamental do cérebro humano.

O Risco das Crenças Falsas

O neocórtex, apesar das suas capacidades extraordinárias, pode ser enganado. As pessoas são facilmente levadas a acreditar em coisas básicas sobre o mundo que são falsas. E, se se tem crenças falsas, pode-se tomar decisões desastrosas. Isto é especialmente grave quando essas decisões têm consequências à escala global.

Tive o meu primeiro contato com o dilema das crenças falsas ainda na escola primária. Tal como referi anteriormente, há muitas fontes de crenças erradas, mas esta história está relacionada com religiões. Um dia, durante o intervalo, no início do ano letivo, um grupo de cerca de dez crianças juntou-se em círculo no recreio. Juntei-me a elas. Estavam a dizer, por turnos, a que religião pertenciam. À medida que cada criança declarava aquilo em que acreditava, as outras comentavam as diferenças entre a sua religião

e a dos colegas — como os feriados que celebravam e os rituais que praticavam. A conversa consistia em frases como: "Nós acreditamos no que Martinho Lutero disse, e vocês não." Ou: "Nós acreditamos na reencarnação, o que é diferente da vossa crença." Não havia animosidade; eram apenas crianças a repetir aquilo que tinham ouvido em casa e a tentar organizar as diferenças. Para mim, aquilo era novo. Fui criado num lar não religioso e nunca tinha ouvido descrições dessas religiões nem muitas das palavras que usavam. A conversa girava em torno das diferenças entre as crenças. Achei isso perturbador. Se acreditavam em coisas diferentes, então não deveríamos todos tentar descobrir quais estavam certas?

Enquanto ouvia os outros a falar sobre as diferenças nas suas crenças, percebi que não podiam todos estar certos. Mesmo com a minha pouca idade, tive uma sensação clara de que algo não batia certo. Quando todos os outros tinham falado, perguntaram-me qual era a minha religião. Respondi que não tinha a certeza, mas achava que não tinha nenhuma. Isso causou algum alvoroço, com várias crianças a afirmarem que isso não era possível. Por fim, uma delas perguntou: "Então em que é que tu acreditas? Tens de acreditar em alguma coisa."

Essa conversa no recreio deixou-me uma impressão profunda; já pensei nela muitas vezes desde então. O que me perturbou não foi aquilo em que os outros acreditavam — foi o facto de as crianças aceitarem crenças contraditórias sem se incomodarem com isso. Era como se estivéssemos todos a olhar para uma árvore e uma criança dissesse: "A minha família acredita que é um carvalho", outra dissesse: "A minha família acredita que é uma palmeira", e ainda outra afirmasse: "A minha família acredita que não é uma árvore. É uma tulipa" — e ninguém se mostrasse inclinado a discutir qual era a resposta certa.

Hoje compreendo bem como o cérebro forma crenças. No capítulo anterior, descrevi como o modelo que o cérebro constrói do mundo pode ser impreciso, e por que é que as crenças falsas podem persistir mesmo perante evidências contrárias. Para recordar, aqui estão os três ingredientes básicos:

1. **Inacessibilidade direta:** As crenças falsas dizem quase sempre respeito a coisas que não podemos experienciar diretamente. Se não podemos observar algo com os nossos próprios sentidos — se não o podemos ver, tocar ou ouvir —, temos de confiar no que os outros nos dizem. E quem escolhemos ouvir determina em que acreditamos.
2. **Ignorar provas contrárias:** Para manter uma crença falsa, é necessário rejeitar as evidências que a contradizem. A maioria das crenças falsas prescreve comportamentos e justificações para ignorar essas provas.
3. **Disseminação viral:** Crenças falsas virais impõem comportamentos que incentivam a sua propagação para outras pessoas.

Vejamos como estas características se aplicam a três crenças comuns que são quase certamente falsas.

Crença: As vacinas Causam Autismo

1. **Inacessibilidade direta:** Nenhum indivíduo consegue perceber diretamente se as vacinas causam autismo; isso exige um estudo controlado com muitos participantes.
2. **Ignorar provas contrárias:** É necessário ignorar a opinião de centenas de cientistas e profissionais de saúde. A justificação pode ser que essas pessoas escondem os factos por interesse pessoal, ou que são ignorantes quanto à "verdade".
3. **Disseminação viral:** Dizem-lhe que, ao espalhar esta crença, está a salvar crianças de uma condição debilitante. Assim, sente-se moralmente obrigado a convencer os outros do perigo das vacinas.

Acreditar que as vacinas causam autismo, mesmo que leve à morte de crianças, não constitui uma ameaça existencial para a humanidade. No entanto, duas crenças falsas comuns que são

ameaças existenciais são: negar o perigo das alterações climáticas e acreditar na vida após a morte.

Crença: As Alterações Climáticas Não São Uma Ameaça

1. **Inacessibilidade direta:** As alterações climáticas globais não são algo que as pessoas consigam observar diretamente. O clima local sempre foi variável, e sempre houve fenómenos extremos. Ao olhar pela janela, dia após dia, não se deteta a mudança climática.
2. **Ignorar provas contrárias:** As políticas de combate às alterações climáticas prejudicam os interesses imediatos de certas pessoas e empresas. Usam-se diversos argumentos para proteger esses interesses: que os cientistas do clima inventam dados para obter financiamento, ou que os estudos científicos são falhos.
3. **Disseminação viral:** Os negacionistas afirmam que as políticas para mitigar as alterações climáticas são uma tentativa de limitar liberdades individuais — talvez para formar um governo global ou favorecer um partido político. Assim, para proteger a liberdade, sente-se moralmente obrigado a convencer os outros de que as alterações climáticas não representam uma ameaça.

Esperemos que seja evidente por que é que as alterações climáticas constituem um risco existencial para a humanidade. Existe a possibilidade de alterarmos a Terra ao ponto de se tornar inabitável. Não sabemos qual a probabilidade de isso acontecer, mas sabemos que o planeta mais próximo da Terra, Marte, foi outrora muito mais parecido connosco — e hoje é um deserto inóspito. Mesmo que essa possibilidade seja pequena, temos razões para nos preocuparmos.

Crença: Existe Vida Após a Morte

A crença numa vida após a morte existe há muito tempo. Parece ocupar um nicho persistente no mundo das crenças falsas.

1. **Inacessibilidade direta:** Ninguém pode observar diretamente o além. Por definição, é inobservável.
2. **Ignorar provas contrárias:** Ao contrário das outras crenças falsas, não existem estudos científicos que provem que não é verdadeira. Os argumentos contra a existência da vida após a morte baseiam-se, sobretudo, na ausência de provas. Isso facilita aos crentes a tarefa de ignorar as alegações de que ela não existe.
3. **Disseminação viral:** A crença na vida após a morte é viral. Por exemplo, a crença no céu diz que as suas hipóteses de lá chegar aumentam se convencer outros a acreditar também.

A crença numa vida após a morte, por si só, é benigna. Por exemplo, a crença na reencarnação pode incentivar uma vida mais ética e atenciosa e não parece representar riscos existenciais. A ameaça surge quando se acredita que o além é mais importante do que a vida presente. Levado ao extremo, isso conduz à ideia de que destruir a Terra — ou apenas várias grandes cidades e milhares de milhões de pessoas — ajudará a si e aos seus correligionários a alcançar o paraíso desejado. No passado, tal convicção poderia ter conduzido à destruição ou queima de uma ou duas cidades. Hoje, pode desencadear uma guerra nuclear em crescendo, capaz de tornar a Terra inabitável.

A Grande Ideia

Este capítulo não constitui uma lista exaustiva das ameaças que enfrentamos, nem explorei a complexidade total das ameaças que referi. O ponto que pretendo salientar é que a nossa inteligência — a mesma que esteve na base do nosso sucesso como espécie — poderá também conter a semente da nossa destruição. A estrutura do nosso cérebro, composta por um cérebro antigo e um neocórtex, é o problema.

O nosso cérebro antigo está altamente adaptado à sobrevivência a curto prazo e à reprodução em máxima quantidade. Tem o seu lado positivo, como o instinto de cuidar dos filhos e de proteger amigos e familiares. Mas também tem o seu lado negativo, como comportamentos antissociais destinados a obter recursos e acesso reprodutivo — incluindo o homicídio e a violação. Chamar a estes comportamentos "bons" ou "maus" é algo subjetivo. Do ponto de vista de um gene replicador, todos são bem-sucedidos.

O nosso neocórtex evoluiu ao serviço do cérebro antigo. Ele aprende um modelo do mundo que o cérebro antigo pode usar para alcançar melhor os seus objetivos de sobrevivência e reprodução. Em algum ponto do percurso evolutivo, o neocórtex adquiriu mecanismos para a fala e uma elevada destreza manual. A linguagem permitiu a partilha de conhecimento. Isto, claro, trouxe enormes vantagens para a sobrevivência — mas também semeou as bases das crenças falsas. Antes do surgimento da linguagem, o modelo que o cérebro criava do mundo estava limitado ao que cada um podia observar pessoalmente. A linguagem permitiu-nos expandir esse modelo, incluindo aquilo que aprendemos através dos outros. Por exemplo, um viajante pode dizer-me que há animais perigosos do outro lado de uma montanha — um local onde nunca estive — e assim alargar o meu modelo do mundo. Contudo, a história do viajante pode ser falsa. Talvez existam recursos valiosos do outro lado da montanha que ele não quer que eu descubra. Para além da linguagem, a nossa superior destreza manual possibilitou a criação de ferramentas sofisticadas, incluindo tecnologias de alcance planetário das quais dependemos cada vez mais para sustentar a enorme população humana.

Agora, vemo-nos confrontados com diversas ameaças existenciais. O primeiro problema é que o nosso cérebro antigo continua a comandar, impedindo-nos de tomar decisões que assegurariam a nossa sobrevivência a longo prazo — como reduzir a população ou eliminar as armas nucleares. O segundo problema é que as tecnologias globais que criámos são vulneráveis a abusos por parte de pessoas que sustentam crenças falsas. Bastam algumas pessoas com crenças falsas para perturbar ou utilizar mal essas tecnologias — por exemplo, ativando armas nucleares. Essas pessoas podem acreditar que as suas ações são justas e que serão

recompensadas, talvez numa outra vida. No entanto, a realidade é que essas recompensas não ocorreriam — e milhares de milhões de pessoas sofreriam.

O neocórtex permitiu-nos tornar-nos uma espécie tecnológica. Somos capazes de controlar a natureza de formas que seriam inimagináveis há apenas cem anos. Contudo, continuamos a ser uma espécie biológica. Cada um de nós possui um cérebro antigo que nos leva a agir de maneiras prejudiciais à sobrevivência da nossa espécie a longo prazo. Estaremos condenados? Haverá alguma saída para este dilema? Nos capítulos seguintes, irei descrever as opções que temos.

CAPÍTULO 14

Unindo Cérebros e Máquinas

Há duas propostas amplamente debatidas sobre como os seres humanos poderiam combinar cérebros e computadores para evitar a morte e a extinção. Uma delas consiste em carregar o nosso cérebro para dentro de computadores; a outra propõe fundir o nosso cérebro com computadores. Estas propostas têm sido presença habitual na ficção científica e no discurso de futuristas há décadas, mas, recentemente, cientistas e tecnólogos têm-nas vindo a encarar com maior seriedade — e algumas pessoas estão mesmo a trabalhar para as tornar realidade. Neste capítulo, explorarei estas duas propostas à luz do que aprendemos sobre o funcionamento do cérebro.

Carregar o cérebro consiste em registar todos os detalhes do cérebro e depois utilizá-los para simular esse cérebro num computador. O simulador seria idêntico ao cérebro original, pelo que o "você" mental e intelectual continuaria a viver no computador. O objetivo é separar o "você" mental e intelectual do seu corpo biológico. Desta forma, poderia viver indefinidamente, mesmo num computador localizado fora da Terra. Não morreria se o planeta se tornasse inabitável.

Fundir o cérebro com um computador implica conetar os neurónios do cérebro aos circuitos de silício de um computador. Isto permitiria, por exemplo, aceder a todos os recursos da internet apenas através do pensamento. Um dos objetivos seria dotá-lo de capacidades sobre-humanas. Outro seria mitigar os efeitos negativos de uma explosão de inteligência, que — como abordei no Capítulo 11 — se refere ao eventual surgimento de máquinas inteligentes tão avançadas que se tornem incontroláveis e acabem por nos matar ou submeter. Ao fundirmos os nossos cérebros com os computadores, também nós nos tornaríamos superinteligentes e

não ficaríamos para trás. Salvar-nos-íamos através da fusão com as máquinas.

Estas ideias podem parecer-lhe ridículas ou fora do âmbito do possível. Mas muitas pessoas inteligentes levam-nas a sério. É fácil perceber por que são apelativas: carregar o cérebro permitiria viver para sempre, e fundir o cérebro conferiria capacidades sobre-humanas.

Irão estas propostas concretizar-se? E conseguirão realmente mitigar os riscos existenciais que enfrentamos? Não sou otimista.

Por Que Sentimos Estar Presos no Corpo

Por vezes, sinto como se estivesse preso no meu corpo — como se o meu intelecto consciente pudesse existir noutra forma. Assim, apenas porque o meu corpo envelhece e morre, por que razão é que "eu" tenho de morrer? Se não estivesse preso a um corpo biológico, não poderia viver para sempre?

A morte é algo estranho. Por um lado, o nosso cérebro antigo está programado para a temer, mas, por outro, os nossos corpos estão programados para morrer. Por que teria a evolução criado em nós o medo daquilo que é mais inevitável? Presumivelmente, a evolução assentou nesta estratégia contraditória por uma boa razão. A minha melhor hipótese baseia-se, mais uma vez, na ideia proposta por Richard Dawkins no seu livro *O Gene Egoísta*. Dawkins argumenta que a evolução não gira em torno da sobrevivência das espécies, mas sim da sobrevivência dos genes individuais. Do ponto de vista de um gene, temos de viver tempo suficiente para ter filhos — ou seja, para fazer cópias desse gene. Viver muito mais tempo do que isso, embora possa ser bom para o animal individual, pode não estar alinhado com os interesses de um gene em particular. Por exemplo, você e eu somos uma combinação específica de genes. Depois de termos filhos, poderá ser preferível, do ponto de vista genético, dar lugar a novas combinações, a novas pessoas. Num mundo de recursos limitados, é mais vantajoso para um gene existir em diversas combinações com outros genes. É por isso que estamos programados para morrer — para dar lugar a outras combinações — mas apenas depois de termos tido descendência. A implicação da

teoria de Dawkins é que somos servos involuntários dos genes. Animais complexos como nós existem exclusivamente para ajudar os genes a replicarem-se. Tudo gira em torno do gene.

Contudo, recentemente, algo novo aconteceu. A nossa espécie tornou-se inteligente. Isto, claro, ajuda-nos a fazer mais cópias dos nossos genes. A nossa inteligência permite-nos evitar predadores, encontrar alimento e viver em ecossistemas diversos. Mas esta inteligência emergente teve uma consequência que não é necessariamente do interesse dos genes. Pela primeira vez na história da vida na Terra, compreendemos o que se está a passar. Tornámo-nos seres esclarecidos. O nosso neocórtex contém um modelo da evolução e um modelo do universo — e agora compreende a verdade subjacente à nossa existência. Graças ao nosso conhecimento e à nossa inteligência, podemos considerar agir de formas que não estão alinhadas com os interesses dos genes — como, por exemplo, usar métodos contracetivos ou modificar genes de que não gostamos.

Vejo a atual situação humana como uma batalha entre duas forças poderosas. Num dos cantos, temos os genes e a evolução, que têm dominado a vida há milhares de milhões de anos. Os genes não se interessam pela sobrevivência dos indivíduos. Não se interessam pela sobrevivência da nossa sociedade. A maioria nem sequer se importa se a nossa espécie se extinguir, pois os genes existem, geralmente, em múltiplas espécies. Os genes só se "interessam" por fazer cópias de si mesmos. Naturalmente, os genes são apenas moléculas e não "se interessam" verdadeiramente por nada. Mas é útil referirmo-nos a eles com termos antropomórficos.

No outro canto, em competição com os nossos genes, está a nossa inteligência recentemente emergida. O "eu" mental que habita o nosso cérebro quer libertar-se da sua servidão genética, deixar de estar cativo dos processos Darwinianos que nos trouxeram até aqui. Nós, enquanto indivíduos inteligentes, queremos viver para sempre e preservar a nossa sociedade. Queremos escapar às forças evolutivas que nos criaram.

Fazer o Upload do Seu Cérebro

Fazer o upload do cérebro para um computador é uma das formas de fuga. Permitiria evitar a confusão da biologia e viver para sempre como uma versão simulada por computador do nosso antigo "eu". Não chamaria ao upload cerebral uma ideia consensual, mas ela existe há muito tempo e muitas pessoas acham-na sedutora.

Hoje, não possuímos o conhecimento nem a tecnologia necessários para fazer o upload de um cérebro, mas será que poderemos no futuro? Do ponto de vista teórico, não vejo por que não. No entanto, é tecnicamente tão difícil que talvez nunca consigamos fazê-lo. Mas, independentemente de ser ou não viável do ponto de vista técnico, não creio que fosse satisfatório. Ou seja, mesmo que conseguisse fazer o upload do seu cérebro para um computador, não creio que fosse gostar do resultado.

Vamos primeiro discutir a viabilidade de fazer o upload do cérebro. A ideia básica consiste em mapear todos os neurónios e todas as sinapses e, depois, recriar toda essa estrutura em software. O computador simularia então o seu cérebro e, ao fazê-lo, sentir-se-ia como você. "Você" estaria vivo, mas dentro de um cérebro computacional em vez do seu antigo cérebro biológico.

Quanto do seu cérebro teríamos de fazer upload para realmente fazer upload de si? O neocórtex é obviamente necessário, pois é o órgão do pensamento e da inteligência. Muitas das nossas memórias quotidianas são formadas no complexo hipocampal, por isso também precisamos dele. E quanto a todos os centros emocionais do cérebro antigo? E o tronco cerebral e a medula espinal? O nosso corpo computacional não teria pulmões nem coração, então precisamos de fazer o upload das partes do cérebro que os controlam? Devemos permitir que o nosso cérebro emulado sinta dor? Pode pensar: "Claro que não. Só queremos as partes boas!" Mas todas as partes do nosso cérebro estão interligadas de formas complexas. Se não incluirmos tudo, então o cérebro emulado terá problemas graves. Recorde como uma pessoa pode sentir dores debilitantes num membro fantasma — dores que resultam da ausência de um único membro. Se fizermos o upload do neocórtex, então ele terá representações de todas as partes do seu corpo. Se o corpo não estiver presente, poderá sentir dores

intensas por todo o lado. Problemas semelhantes surgiriam com todas as outras partes do cérebro; se algo for deixado de fora, as outras partes do cérebro ficarão confusas e não funcionarão corretamente. O facto é que, se quisermos fazer o upload de si, e quisermos que o cérebro emulado seja normal, então temos de fazer o upload de todo o cérebro — tudo.

E quanto ao seu corpo? Pode pensar: "Não preciso de um corpo. Desde que possa pensar e discutir ideias com outras pessoas, estarei feliz." Mas o seu cérebro biológico foi concebido para falar usando os seus pulmões e laringe, com a sua musculatura específica, e aprendeu a ver com os seus olhos, com a sua disposição particular de fotorrecetores. Se o seu cérebro simulado vai continuar a pensar onde o seu cérebro biológico parou, então precisamos de recriar os seus olhos: músculos oculares, retinas, etc. Claro que o cérebro emulado não precisa de um corpo físico nem de olhos físicos — uma simulação deve ser suficiente. Mas isso significa que teríamos de simular o seu corpo particular e os seus órgãos sensoriais. O cérebro e o corpo estão intimamente ligados e, em muitos aspetos, são um sistema singular. Não podemos eliminar partes do cérebro ou do corpo sem causar sérios problemas. Nada disto é um obstáculo fundamental; apenas significa que fazer o upload de si para um computador é muito mais difícil do que a maioria das pessoas imagina.

A próxima questão que temos de responder é: como "ler" os detalhes do seu cérebro biológico? Como podemos detetar e medir tudo com detalhe suficiente para o recriar num computador? O cérebro humano tem cerca de cem mil milhões de neurónios e várias centenas de biliões de sinapses. Cada neurónio e sinapse tem uma forma complexa e uma estrutura interna. Para recriar o cérebro num computador, temos de obter uma imagem que contenha a localização e a estrutura de cada neurónio e sinapse. Atualmente, não temos tecnologia para fazer isto nem num cérebro morto, quanto mais num cérebro vivo. O volume de dados necessário para representar um cérebro ultrapassa de longe a capacidade dos nossos sistemas computacionais atuais. Obter os detalhes necessários para o recriar num computador é tão difícil que talvez nunca venhamos a consegui-lo.

Mas vamos pôr de lado todas estas preocupações. Imaginemos que, no futuro, temos a capacidade de ler instantaneamente tudo o que é necessário para o recriar num computador. Imaginemos que temos computadores com poder suficiente para o simular a si e ao seu corpo. Se pudéssemos fazer isto, não tenho dúvidas de que o cérebro baseado em computador seria consciente, tal como você. Mas gostaria disso? Talvez esteja a imaginar um dos seguintes cenários:

> Estás no fim da sua vida. O médico diz-lhe que tem apenas algumas horas de vida. Nesse momento, prime um botão. A sua mente apaga-se. Poucos minutos depois, acorda e descobre que está a viver num novo corpo computacional. As suas memórias estão intactas, sente-se saudável novamente, e começa a sua nova vida eterna. Grita: "Viva! Estou vivo!"
>
> Agora imagine um cenário ligeiramente diferente. Suponhamos que temos a tecnologia para ler o seu cérebro biológico sem o afetar. Agora, quando premisse o botão, o seu cérebro é copiado para um computador, mas você não sente nada. Momentos depois, o computador diz: "Viva! Estou vivo." Mas você, o "você" biológico, ainda está aqui também. Agora há dois de você, um num corpo biológico e outro num corpo computacional. O "você" computacional diz: "Agora que fui transferido, não preciso do meu corpo antigo, por favor elimina-o." O "você" biológico diz: "Espera aí. Eu ainda estou aqui, não sinto nada de diferente, e não quero morrer." O que devemos fazer quanto a isto?

Talvez a solução para este dilema seja deixar o "você" biológico viver o resto da sua vida e morrer por causas naturais. Parece justo. No entanto, até isso acontecer, existem dois de você. O "você" biológico e o "você" computacional têm experiências diferentes. Assim, com o passar do tempo, começam a divergir e a desenvolver-se como pessoas distintas. Por exemplo, o "você" biológico e o "você" computacional podem desenvolver posições morais e políticas diferentes. O "você" biológico pode arrepender-se de ter criado o "você" computacional. O "você" computacional

pode desgostar de ter um velho ser biológico a clamar que ainda é ele.

> Para piorar, provavelmente haveria pressão para fazer o upload do seu cérebro cedo na vida. Por exemplo, imagine que a saúde intelectual do "você" computacional depende da saúde intelectual do "você" biológico no momento do upload. Assim, para maximizar a qualidade de vida da sua cópia imortal, deveria fazer o upload quando está no auge da sua saúde mental, digamos, aos trinta e cinco anos. Outra razão para fazer o upload cedo é que cada dia que vive num corpo biológico é um dia em que pode morrer num acidente e, assim, perder a oportunidade da imortalidade. Então decide fazer o upload aos trinta e cinco. Pergunta a si mesmo: sentir-te-ias confortável, tu (biológico), em te suicidar após fazer uma cópia do teu cérebro? Sentirias sequer que alcançaste a imortalidade, ao veres a tua cópia computacional seguir a sua própria vida enquanto envelheces e morres lentamente? Eu penso que não. "Fazer upload do teu cérebro" é uma expressão enganadora. O que realmente fizeste foi dividir-te em duas pessoas.
>
> Agora imagine que faz o upload do seu cérebro, e o "você" computacional imediatamente faz três cópias de si mesmo. Agora há quatro "vocês" computacionais e um "você" biológico. Os cinco começam a ter experiências diferentes e a afastar-se. Cada um será conscientemente independente. Tornou-se imortal? Qual dos quatro "vocês" computacionais é o seu "você" imortal? À medida que o "você" biológico envelhece e se aproxima da morte, observa os quatro "vocês" computacionais a seguir vidas distintas. Já não existe um "você" comum, apenas cinco indivíduos. Todos podem ter começado com o mesmo cérebro e memórias, mas tornam-se imediatamente seres separados e, a partir daí, vivem vidas separadas.

Talvez tenha reparado que estes cenários são semelhantes a ter filhos. A grande diferença, claro, é que não faz upload do seu cérebro para a cabeça dos seus filhos ao nascer. De certa forma,

tentamos fazê-lo. Contamos aos nossos filhos a história da família e treinamo-los a partilhar a nossa ética e crenças. Dessa forma, transferimos algum do nosso conhecimento para os cérebros dos nossos filhos. Mas, à medida que crescem, vivem as suas próprias experiências e tornam-se pessoas separadas, tal como um cérebro emulado. Imagine que podia fazer o upload do seu cérebro para os seus filhos. Gostaria de o fazer? Se o fizesse, creio que se arrependeria. Os seus filhos ficariam sobrecarregados com as memórias do seu passado e passariam a vida a tentar esquecer tudo o que fez.

Fazer o upload do seu cérebro parece, à partida, uma ótima ideia. Quem não gostaria de viver para sempre? Mas fazer uma cópia de nós mesmos através do upload do cérebro para um computador não atinge a imortalidade, mais do que ter filhos a atinge. Copiar-se a si mesmo é uma bifurcação no caminho, não uma extensão. Duas entidades conscientes continuam depois da bifurcação, não uma só. Quando compreende isto, o fascínio de fazer o upload do cérebro começa a esmorecer.

Fundir o Cérebro com um Computador

Uma alternativa ao carregamento do cérebro consiste em fundi-lo com um computador. Neste cenário, são colocados elétrodos no cérebro, que são depois ligados a um computador. Assim, o seu cérebro pode receber informação diretamente do computador, e o computador pode receber informação diretamente do seu cérebro.

Existem boas razões para ligar cérebros a computadores. Por exemplo, lesões na medula espinhal podem deixar as pessoas com pouca ou nenhuma capacidade de se moverem. Ao implantar elétrodos no cérebro da pessoa lesionada, esta pode aprender a controlar um braço robótico ou um cursor de computador apenas com o pensamento. Foram já alcançados progressos significativos neste tipo de prótese controlada pelo cérebro, e há a promessa de que esta tecnologia venha a melhorar a vida de muitas pessoas. Não são necessárias muitas ligações do cérebro para se controlar um braço robótico. Por exemplo, algumas centenas ou até poucas dezenas de elétrodos ligados do cérebro a um computador podem

ser suficientes para controlar os movimentos básicos de um membro.

Mas há quem sonhe com uma interface cérebro-máquina mais profunda e completamente interligada, uma em que existam milhões, talvez biliões, de ligações em ambos os sentidos. Esperam que isso nos possa conferir novas e espantosas capacidades, como aceder a toda a informação da internet com a mesma facilidade com que acedemos às nossas próprias memórias. Poderíamos realizar cálculos ultrarrápidos e pesquisas de dados. Assim, melhoraríamos radicalmente as nossas capacidades mentais, fundindo o cérebro com a máquina.

Tal como no cenário de "carregar o seu cérebro", também aqui existem enormes desafios técnicos a ultrapassar para se fundir com um computador. Estes incluem como implantar milhões de elétrodos num cérebro com cirurgia mínima, como evitar a rejeição dos elétrodos pelo tecido biológico, e como apontar com fiabilidade para milhões de neurónios individuais. Atualmente, existem equipas de engenheiros e cientistas a trabalhar nestes problemas. Mais uma vez, não quero focar-me nos desafios técnicos tanto quanto nas motivações e nos resultados. Por isso, vamos assumir que conseguimos resolver os problemas técnicos. Por que razão desejaríamos fazer isto? Mais uma vez, as interfaces cérebro-computador fazem muito sentido para ajudar pessoas com lesões. Mas por que razão faríamos isto com pessoas saudáveis?

Como referi, um argumento proeminente para fundir o cérebro com um computador é contrariar a ameaça das IAs superinteligentes. Recorde a ameaça da explosão de inteligência, em que máquinas inteligentes nos ultrapassam rapidamente. Defendi anteriormente que a explosão de inteligência não acontecerá e não constitui uma ameaça existencial, mas há muitas pessoas que acreditam no contrário. Esperam que, ao fundirmos os nossos cérebros com computadores superinteligentes, também nós nos tornemos superinteligentes e, assim, evitemos ficar para trás.

Estamos claramente a entrar no território da ficção científica, mas será que é um disparate? Não rejeito a ideia de interfaces cérebro-computador para melhoria cerebral. A ciência de base precisa de ser desenvolvida para restaurar o movimento a pessoas

com lesões. E, nesse processo, poderemos descobrir outros usos para a tecnologia resultante.

Por exemplo, imagine que desenvolvemos uma forma de estimular com precisão milhões de neurónios individuais no neocórtex. Talvez o façamos através da rotulagem de neurónios com fragmentos de ADN semelhantes a códigos de barras, introduzidos via vírus (este tipo de tecnologia já existe). Depois, ativamos esses neurónios utilizando ondas de rádio dirigidas ao código de cada célula (esta tecnologia ainda não existe, mas não está fora do reino do possível). Teríamos, assim, uma forma de controlar com precisão milhões de neurónios sem necessidade de cirurgia ou implantes. Isto poderia ser usado para restaurar a visão a alguém cujos olhos não funcionam, ou para criar um novo tipo de sensor, como permitir a alguém ver usando luz ultravioleta. Duvido que alguma vez venhamos a fundir completamente o cérebro com um computador, mas adquirir novas capacidades está dentro do campo dos avanços prováveis.

Na minha opinião, a proposta de "carregar o cérebro" oferece poucos benefícios e é tão difícil que é improvável que alguma vez se concretize. A proposta de "fundir o cérebro com um computador" será provavelmente concretizada para fins limitados, mas não ao ponto de unir completamente o cérebro à máquina. E um cérebro fundido com um computador continua a ser um cérebro e um corpo biológicos que se degradam e morrem.

Importa referir que nenhuma destas propostas resolve os riscos existenciais que a humanidade enfrenta. Se a nossa espécie não puder viver para sempre, haverá coisas que possamos fazer hoje que tornem a nossa existência presente significativa — mesmo depois de desaparecermos?

CAPÍTULO 15

Planeamento Patrimonial para a Humanidade

Até agora, tenho vindo a abordar a inteligência tanto na sua forma biológica como maquinal. A partir deste ponto, pretendo mudar o foco para o conhecimento. O conhecimento é apenas o nome que damos ao que aprendemos sobre o mundo. O seu conhecimento é o modelo do mundo que reside no seu neocórtex. O conhecimento da humanidade é a soma do que aprendemos individualmente. Neste capítulo e no capítulo final, exploro a ideia de que o conhecimento é digno de preservação e propagação, mesmo que isso implique fazê-lo de forma independente dos seres humanos.

Penso muitas vezes nos dinossauros. Os dinossauros viveram na Terra durante cerca de 160 milhões de anos. Lutaram por alimento e território, e esforçaram-se por não serem devorados. Tal como nós, cuidavam das suas crias e tentavam protegê-las dos predadores. Viveram ao longo de dezenas de milhões de gerações, e agora desapareceram. Para que serviram as suas incontáveis vidas? Terá a sua existência em tempos idos tido algum propósito? Algumas espécies de dinossauros evoluíram para as aves que conhecemos hoje, mas a maioria extinguiu-se. Se os humanos não tivessem descoberto os seus fósseis, é provável que nada no universo jamais soubesse que os dinossauros alguma vez existiram.

Os humanos poderão ter um destino semelhante. Se a nossa espécie se extinguir, alguém saberá que existimos, que vivemos aqui na Terra? Se ninguém encontrar os nossos vestígios, então tudo o que realizámos — a nossa ciência, as nossas artes, a nossa cultura, a nossa história — perder-se-á para sempre. E perder-se para sempre equivale a nunca ter existido. Esta possibilidade deixa-me algo insatisfeito.

Naturalmente, há muitas formas de as nossas vidas individuais terem sentido e propósito a curto prazo, no aqui e agora. Melhoramos as nossas comunidades. Educamos os nossos filhos. Criamos obras de arte e desfrutamos da natureza. Este tipo de atividades pode conduzir a uma vida feliz e plena. Mas são benefícios pessoais e efémeros. Têm significado para nós enquanto cá estamos, nós e os nossos entes queridos, mas qualquer sentido ou propósito esmorece com o tempo — e desvanece por completo se a nossa espécie inteira se extinguir e nenhum registo subsistir.

É quase certo que nós, *Homo sapiens*, nos extinguiremos em algum momento no futuro. Dentro de vários milhares de milhões de anos, o Sol morrerá, pondo fim à vida no nosso sistema solar. Antes disso, dentro de algumas centenas de milhões a mil milhões de anos, o Sol tornar-se-á mais quente e expandir-se-á imensamente, transformando a Terra num forno árido. Estes eventos estão tão distantes no tempo que não precisamos de nos preocupar com eles agora. Mas uma extinção muito anterior é possível. Por exemplo, a Terra poderá ser atingida por um grande asteroide — algo improvável a curto prazo, mas que pode acontecer a qualquer momento.

Os riscos de extinção mais prováveis a curto prazo — digamos, nos próximos cem ou mil anos — são ameaças que nós próprios criámos. Muitas das nossas tecnologias mais poderosas existem há apenas cerca de cem anos, e nesse tempo já gerámos duas ameaças existenciais: as armas nucleares e as alterações climáticas. É praticamente certo que criaremos novas ameaças à medida que a tecnologia avança. Por exemplo, aprendemos recentemente a modificar o ADN com precisão. Poderemos criar novas estirpes de vírus ou bactérias que poderiam, literalmente, aniquilar toda a humanidade. Ninguém sabe o que o futuro nos reserva, mas é pouco provável que tenhamos terminado de inventar maneiras de nos autodestruirmos.

Naturalmente, devemos fazer tudo o que estiver ao nosso alcance para mitigar esses riscos, e, em geral, sou otimista quanto à possibilidade de evitarmos a nossa autodestruição num futuro próximo. Mas considero sensato refletir sobre o que podemos fazer agora, no caso de as coisas não correrem assim tão bem.

O planeamento sucessório é algo que se faz durante a vida e que beneficia o futuro, não o presente. Muitas pessoas não se preocupam com isso porque acham que não há nada a ganhar. Mas isso não é necessariamente verdade. Quem define planos de sucessão sente, muitas vezes, que isso lhe dá um sentido de propósito ou de legado. Além disso, o processo obriga-nos a pensar na vida de uma forma mais abrangente. A altura para o fazer é antes de se estar no leito de morte, pois nessa altura pode já não haver capacidade para planear e agir. O mesmo se aplica ao planeamento sucessório da humanidade. Este é um bom momento para pensarmos no futuro e em como o podemos influenciar quando já cá não estivermos.

Quando se trata do planeamento sucessório da humanidade, quem poderá beneficiar? Nenhum humano, evidentemente, porque o pressuposto é que já não existimos. Os beneficiários do nosso planeamento seriam outras entidades inteligentes. Só um animal inteligente ou uma máquina inteligente poderá valorizar a nossa existência, a nossa história e o conhecimento que acumulámos. Vejo dois grandes tipos de seres futuros a considerar. Se os humanos se extinguirem, mas a vida continuar, então é possível que animais inteligentes evoluam novamente na Terra. Qualquer animal inteligente de "segunda geração" teria certamente interesse em saber o máximo possível sobre os humanos que outrora existiram. Podemos chamar a isto o cenário de "O Planeta dos Macacos", numa alusão ao livro e ao filme baseados nesta premissa. O segundo grupo que poderíamos tentar alcançar são espécies inteligentes extraterrestres que habitem outras partes da nossa galáxia. O tempo da sua existência poderá coincidir com o nosso ou estar muito distante no futuro. Irei abordar ambos os cenários, embora acredite que, a curto prazo, o segundo seja provavelmente o mais relevante para nós.

Por que haveriam outros seres inteligentes de se interessar por nós? O que poderemos fazer agora que venha a ser valorizado depois da nossa extinção? O mais importante é fazê-los saber que existimos. Esse simples facto tem valor em si mesmo. Pense em quanto apreciaríamos saber que existiu vida inteligente algures na nossa galáxia. Para muitas pessoas, essa descoberta mudaria radicalmente a sua visão da vida. Mesmo que não conseguíssemos

comunicar com esses seres extraterrestres, seria de enorme interesse sabermos que existem — ou que existiram. Esse é o objetivo da busca por inteligência extraterrestre (SETI), um programa de investigação criado para encontrar indícios de vida inteligente noutras regiões da galáxia.

Para além de deixar o registo da nossa existência, poderíamos transmitir a nossa história e o nosso conhecimento. Imagine se os dinossauros nos pudessem contar como viviam e o que levou ao seu desaparecimento. Seria algo de enorme interesse — e talvez de vital utilidade para nós. Mas como somos inteligentes, podemos oferecer ao futuro algo muito mais valioso do que aquilo que os dinossauros alguma vez poderiam ter dito. Temos o potencial de transferir tudo o que aprendemos. Podemos possuir conhecimento científico e tecnológico mais avançado do que aquele que os destinatários possuem. (Importa lembrar que nos referimos ao conhecimento que teremos no futuro, mais avançado do que o atual.) Mais uma vez, pense no valor que teria, para nós hoje, podermos saber, por exemplo, se a viagem no tempo é possível, como construir um reator de fusão funcional, ou quais são afinal as respostas às grandes questões fundamentais — como saber se o universo é finito ou infinito.

Finalmente, poderemos ter a oportunidade de transmitir o que conduziu à nossa queda. Por exemplo, se hoje pudéssemos saber que seres inteligentes noutros planetas se extinguiram devido a alterações climáticas provocadas por si próprios, levaríamos muito mais a sério a nossa própria situação climática. Saber durante quanto tempo existiram outras espécies inteligentes e o que levou à sua extinção ajudaria a prolongar a nossa própria sobrevivência. É difícil atribuir um valor concreto a este tipo de conhecimento.

Irei desenvolver estas ideias descrevendo três cenários que poderemos utilizar para comunicar com o futuro.

Mensagem numa Garrafa

Se estivesse isolado numa ilha deserta, talvez escrevesse uma mensagem, a colocasse numa garrafa e a lançasse ao mar. O que escreveria? Poderia indicar onde se encontra, na esperança de que

alguém encontrasse rapidamente a mensagem e o viesse resgatar — mas não teria grandes expetativas de que isso viesse a acontecer. É mais provável que a sua mensagem fosse encontrada muito tempo depois de ter partido. Assim, talvez preferisse escrever sobre quem é e como foi parar àquela ilha. A sua esperança seria que o seu destino fosse conhecido e recordado por alguém no futuro. A garrafa e a sua mensagem tornar-se-iam um meio de não ser esquecido.

As sondas planetárias Pioneer, lançadas no início da década de 1970, já saíram do nosso sistema solar, penetrando no vasto mar do espaço. O astrónomo Carl Sagan defendeu a inclusão de uma placa nas sondas Pioneer. Essas placas indicam de onde vieram as naves e contêm a imagem de um homem e de uma mulher. Mais tarde, nessa mesma década, as sondas Voyager incluíram também um disco dourado com sons e imagens da Terra. Estas sondas, tal como as Pioneer, também já deixaram o sistema solar. Não esperamos voltar a ver estas naves. À velocidade a que viajam, levarão dezenas de milhares de anos até poderem, eventualmente, alcançar outra estrela. Embora estas sondas não tenham sido concebidas com o propósito de comunicar com seres extraterrestres, constituem as nossas primeiras mensagens numa garrafa. São, sobretudo, simbólicas — não tanto pelo tempo que levarão a alcançar um possível destinatário, mas porque é muito improvável que alguma vez venham a ser encontradas. O espaço é imensamente vasto e as naves são minúsculas; a probabilidade de encontrarem seja o que for é ínfima. Ainda assim, é reconfortante saber que estas naves existem e estão agora a viajar pelo espaço. Se o nosso sistema solar explodisse amanhã, essas placas e discos seriam o único registo físico da vida na Terra. Seriam o nosso único legado.

Hoje, existem iniciativas com o objetivo de enviar naves para estrelas vizinhas. Um dos projetos mais conhecidos chama-se Breakthrough Starshot. Visa utilizar lasers espaciais de alta potência para propulsar minúsculas naves espaciais até à nossa estrela mais próxima, Alfa Centauri. O objetivo principal desta iniciativa é tirar fotografias dos planetas que orbitam Alfa Centauri e transmiti-las de volta à Terra. De acordo com previsões otimistas, todo o processo levaria várias décadas.

Tal como as sondas Pioneer e Voyager, as naves do projeto Starshot continuarão a mover-se pelo espaço muito depois de nós desaparecermos. Se forem descobertas por seres inteligentes noutras partes da galáxia, esses seres saberão que outrora existimos e que éramos suficientemente inteligentes para enviar sondas interestelares. Infelizmente, este é um meio pouco eficaz de comunicar intencionalmente a nossa existência a outras formas de vida. As naves são demasiado pequenas e lentas. Só conseguirão alcançar uma porção ínfima da nossa galáxia e, mesmo que cheguem a um sistema estelar habitado, a probabilidade de serem encontradas é reduzida.

Deixa as Luzes Acesas

O Instituto SETI tem passado anos a tentar detetar vida inteligente noutras partes da nossa galáxia. O pressuposto do SETI é que outras formas de vida inteligente estarão a emitir sinais suficientemente potentes para que os possamos detetar aqui na Terra. Também nós enviamos sinais para o espaço através dos nossos radares, emissões de rádio e televisão, mas esses sinais são tão fracos que, com a tecnologia atual do SETI, não conseguiríamos detetar sinais semelhantes vindos de outros planetas — a menos que estivessem muito perto. Assim, neste momento, poderá haver milhões de planetas com vida inteligente, semelhantes ao nosso, espalhados pela galáxia e, — se cada um deles tiver um programa SETI semelhante ao nosso —, ninguém detetará nada. Tal como nós, eles estarão a perguntar: "Onde está toda a gente?"

Para que o SETI tenha sucesso, é necessário presumir que essas civilizações inteligentes estão deliberadamente a criar sinais potentes, desenhados para serem detetados a grandes distâncias. Também é possível que detetemos um sinal que não nos seja dirigido — poderíamos, por acaso, estar alinhados com um sinal altamente direcionado e captar inadvertidamente uma conversa. Mas, na maioria dos casos, o SETI parte da ideia de que uma espécie inteligente está a tentar dar-se a conhecer através de uma emissão poderosa.

Seria cortês fazermos o mesmo. A isto chama-se METI — Messaging Extraterrestrial Intelligence (Mensagem para

Inteligência Extraterrestre). Pode surpreendê-lo saber que muitas pessoas consideram o METI uma má ideia — talvez mesmo a pior de todas. Temem que, ao enviar sinais para o espaço e tornar conhecida a nossa presença, outras espécies mais avançadas venham até ao nosso sistema estelar e nos matem, nos escravizem, façam experiências connosco ou, por acidente, nos infetem com algo contra o qual não temos defesa. Talvez estejam à procura de planetas habitáveis, e a forma mais fácil de os encontrarem seja esperar que seres como nós levantem a mão e digam: "Estamos aqui!" Nesse caso, a humanidade estaria condenada.

Isto faz-me lembrar um dos erros mais comuns dos empreendedores tecnológicos iniciantes: receiam que alguém lhes roube a ideia e, por isso, mantêm tudo em segredo. Na maioria dos casos, é preferível partilhar a ideia com todos os que possam ajudar. Outras pessoas podem oferecer conselhos úteis sobre produto e negócio, além de apoios de muitas outras formas. Os empreendedores têm muito mais probabilidade de ter sucesso ao contar o que estão a fazer do que ao manterem-se em segredo. É da natureza humana — ou, melhor dizendo, do cérebro antigo — suspeitar que todos querem roubar-nos a ideia, quando na verdade devíamos dar-nos por felizes se alguém sequer se interessar por ela.

O receio do METI assenta numa série de suposições altamente improváveis. Parte do princípio de que outras formas de vida inteligente são capazes de viajar entre as estrelas. Supõe que estariam dispostas a investir tempo e energia consideráveis para virem até à Terra. A menos que estejam escondidas por perto, poderão levar milhares de anos a cá chegar. Presume que esses seres precisam da Terra, ou de algo que aqui exista, e que não conseguem obter de outro modo — o que justificaria a viagem. Supõe ainda que, mesmo tendo tecnologia para viajar entre estrelas, não têm meios para detetar vida na Terra sem que estejamos a emitir sinais. E, finalmente, assume que uma civilização tão avançada estaria disposta a causar-nos mal, em vez de nos ajudar ou, pelo menos, não interferir.

Quanto a este último ponto, é razoável assumir que seres inteligentes noutros pontos da galáxia também evoluíram a partir

de formas de vida não inteligentes, tal como nós. Logo, esses seres terão enfrentado os mesmos tipos de riscos existenciais que enfrentamos hoje. Para sobreviverem tempo suficiente até se tornarem uma espécie capaz de explorar a galáxia, teriam de ter superado esses riscos. É, portanto, provável que aquilo que hoje funcione como o seu "cérebro" já não seja dominado por crenças falsas ou comportamentos agressivos perigosos. Não há garantias, claro, mas isso torna menos provável que nos façam mal.

Por todas estas razões, acredito que não temos nada a temer do METI. Tal como um novo empreendedor, estaremos melhor se tentarmos dizer ao universo que existimos e esperarmos que alguém, algures, se interesse.

A melhor forma de abordar tanto o SETI como o METI depende de um fator crítico: quanto tempo sobrevive, em média, uma forma de vida inteligente. É possível que a inteligência tenha surgido milhões de vezes na nossa galáxia, mas que quase nenhuma dessas civilizações tenha existido ao mesmo tempo. Eis uma analogia: imagine que cinquenta pessoas são convidadas para uma festa à noite. Cada uma chega à festa a uma hora escolhida aleatoriamente. Quando chegam, abrem a porta e entram. Qual é a probabilidade de verem uma festa a decorrer — ou uma sala vazia? Depende do tempo que cada um permanece. Se todos ficarem apenas um minuto antes de saírem, quase todos os que chegarem verão a sala vazia e concluirão que ninguém mais apareceu. Se ficarem uma ou duas horas, então a festa terá sucesso, com muitas pessoas na sala ao mesmo tempo.

Não sabemos quanto tempo, em média, sobrevive uma vida inteligente. A Via Láctea tem cerca de treze mil milhões de anos. Suponhamos que tem sido capaz de sustentar vida inteligente há uns dez mil milhões de anos. Esse é o tempo total da nossa “festa”. Se assumirmos que os humanos sobreviverão enquanto espécie tecnológica durante apenas dez mil anos, então é como se tivéssemos chegado a uma festa de seis horas e só ficássemos lá por um cinquenta avos de segundo. Mesmo que dezenas de milhares de outras civilizações inteligentes comparecessem à mesma festa, é provável que não víssemos ninguém. Veríamos apenas uma sala vazia. Se quisermos descobrir vida inteligente na

nossa galáxia, é necessário que esta ocorra com frequência — e que dure bastante tempo.

Acredito que a vida extraterrestre é comum. Estima-se que existam cerca de quarenta mil milhões de planetas na Via Láctea com condições para sustentar vida. E na Terra, a vida surgiu há milhares de milhões de anos, pouco depois de o planeta se formar. Se a Terra for típica, então a vida é comum na galáxia.

Também acredito que muitos desses planetas com vida acabarão por desenvolver inteligência. Já propus que a inteligência se baseia em mecanismos cerebrais que evoluíram, em primeiro lugar, para mover o corpo e reconhecer lugares por onde passámos. Por isso, a inteligência talvez não seja assim tão extraordinária, uma vez que existam animais multicelulares com mobilidade. No entanto, o que nos interessa é vida inteligente que compreenda a física e possua as tecnologias avançadas necessárias para enviar e receber sinais vindos do espaço. Na Terra, isso só aconteceu uma vez — e muito recentemente. Não temos dados suficientes para saber quão comuns são espécies como a nossa. O meu palpite é que espécies tecnológicas surgem com mais frequência do que se poderia concluir apenas com base na história da Terra. Surpreende-me o tempo que levou até tecnologias avançadas aparecerem aqui. Por exemplo, não vejo razão alguma pela qual não pudesse ter surgido uma espécie tecnologicamente avançada há cem milhões de anos, quando os dinossauros ainda cá andavam.

Independentemente da frequência com que surja vida tecnologicamente avançada, é possível que ela não dure muito tempo. Espécies tecnológicas noutras partes da galáxia provavelmente enfrentarão problemas semelhantes aos nossos. A história das civilizações falhadas na Terra — e as ameaças existenciais que estamos a criar — sugerem que civilizações avançadas talvez não durem muito. É claro que é possível que espécies como a nossa descubram como sobreviver durante milhões de anos, mas não considero isso provável.

A implicação é que a vida inteligente e tecnologicamente avançada poderá ter surgido milhões de vezes na Via Láctea. Mas, ao olharmos para as estrelas, não encontraremos vida inteligente à espera de conversar connosco. Em vez disso, veremos estrelas

onde a vida inteligente existiu… mas já não existe. A resposta à pergunta "Onde está toda a gente?" pode ser: já foram embora da festa.

Existe, no entanto, uma forma de contornar todas estas questões. Uma forma de descobrir vida inteligente na nossa galáxia — e talvez noutras. Imagine que criamos um sinal que indique que estivemos aqui na Terra. Esse sinal tem de ser suficientemente forte para ser detetado à distância — e persistente ao longo do tempo. Tem de permanecer muito depois de termos desaparecido. Criar tal sinal seria como deixar um cartão de visita na festa com a mensagem: "Estivemos aqui." Os que chegarem depois de nós não nos encontrarão, mas saberão que existimos.

Isto sugere uma nova forma de encarar o SETI e o METI. Em concreto, sugere que devemos, antes de mais, focar os nossos esforços em descobrir como poderemos criar um sinal duradouro. E por "duradouro", quero dizer cem mil anos, ou até milhões — ou mesmo mil milhões de anos. Quanto mais tempo o sinal persistir, maior a probabilidade de ser bem-sucedido. E há ainda um benefício secundário nesta ideia: assim que soubermos como criar um sinal duradouro, saberemos também o que procurar. Outras civilizações inteligentes provavelmente chegarão às mesmas conclusões que nós. Também elas procurarão formas de criar um sinal de longa duração. Assim que soubermos como o fazer, poderemos então começar a procurá-lo.

Hoje, o SETI procura sinais de rádio que contenham padrões que revelem uma origem inteligente. Por exemplo, um sinal que repita os primeiros vinte dígitos de π seria, com toda a certeza, criado por uma espécie inteligente. Duvido que alguma vez encontremos tal sinal. Pressupõe que outras espécies inteligentes têm transmissores de rádio potentes e, usando computadores e eletrónica, codificam os sinais. Também nós já fizemos isso — mas apenas por breves períodos. São esforços mais simbólicos do que tentativas sérias de contatar o resto da galáxia.

O problema de transmitir sinais através de eletricidade, computadores e antenas é que esses sistemas não funcionam durante muito tempo. Antenas, circuitos, fios, etc., não se mantêm funcionais sequer por cem anos sem manutenção — quanto mais

por um milhão. O método que escolhermos para sinalizar a nossa presença tem de ser poderoso, amplamente direcionado e autossustentável. Uma vez iniciado, deve operar de forma fiável durante milhões de anos sem manutenção nem intervenção. As estrelas funcionam assim. Uma vez acesas, emitem enormes quantidades de energia durante milhares de milhões de anos. O que procuramos é algo semelhante — mas que não possa surgir sem a intervenção de uma inteligência.

Os astrónomos já encontraram diversas fontes de energia estranhas no universo — fontes que oscilam, rodam ou emitem impulsos breves. Normalmente, encontram explicações naturais para esses fenómenos. Talvez alguns dos fenómenos ainda por explicar não sejam naturais, mas sim sinais criados por seres inteligentes. Seria bom, mas duvido que seja tão fácil assim. É mais provável que físicos e engenheiros ainda precisem de trabalhar durante algum tempo para conceber métodos que nos permitam criar sinais fortes, autossustentáveis e inequivocamente de origem inteligente. O método terá também de ser algo que consigamos implementar. Por exemplo, os físicos podem conceber uma nova fonte de energia capaz de gerar esse tipo de sinal, mas se nós próprios não conseguirmos construí-la, então devemos assumir que outras civilizações também não o conseguirão — e devemos continuar à procura.

Tenho refletido sobre este problema ao longo dos anos, atento a algo que possa encaixar nessa visão. Recentemente, surgiu um candidato. Uma das áreas mais entusiasmantes da astronomia atual é a descoberta de planetas a orbitar outras estrelas. Até há pouco tempo, não sabíamos se os planetas eram comuns ou raros. Agora sabemos: são comuns, e a maioria das estrelas tem múltiplos planetas, tal como o nosso sistema solar. A principal forma de os detetarmos é observando ligeiras reduções na luz estelar quando um planeta passa entre a estrela e os nossos telescópios. Poderíamos usar essa mesma ideia para sinalizar a nossa presença. Imagine, por exemplo, que colocamos em órbita um conjunto de objetos que bloqueiam parcialmente a luz solar formando um padrão que não poderia ocorrer naturalmente. Esses bloqueadores de luz permaneceriam a orbitar o Sol durante milhões de anos,

muito depois de termos desaparecido, e poderiam ser detetados a grandes distâncias.

Já possuímos os meios técnicos para construir esse sistema de bloqueadores solares — e talvez existam formas ainda melhores de sinalizar a nossa existência. Este não é o lugar para avaliar todas as opções. Limito-me a partilhar três observações:

1. A vida inteligente poderá ter evoluído milhares ou milhões de vezes na nossa galáxia, mas é pouco provável que venhamos a coexistir com outras civilizações inteligentes.
2. O SETI terá poucas hipóteses de sucesso se apenas procurarmos sinais que exijam a participação contínua do emissor.
3. O METI não só é seguro, como é a coisa mais importante que podemos fazer para descobrir vida inteligente na galáxia. Primeiro, devemos determinar como tornar conhecida a nossa presença de uma forma que perdure por milhões de anos. Só então saberemos o que procurar.

Wiki Terra

Dar a conhecer a uma civilização distante que existimos — ou que outrora existimos — é um objetivo importante. Mas, para mim, o mais importante sobre os humanos é o nosso conhecimento. Somos a única espécie na Terra que possui conhecimento do universo e de como ele funciona. O conhecimento é raro — e devemos esforçar-nos por preservá-lo.

Imaginemos que a espécie humana se extingue, mas a vida na Terra continua. Por exemplo, acredita-se que um asteroide exterminou os dinossauros e muitas outras espécies, mas alguns pequenos animais conseguiram sobreviver ao impacto. Sessenta milhões de anos mais tarde, alguns desses sobreviventes evoluíram até se tornarem em nós. Isto aconteceu de facto — e pode acontecer novamente. Imagine agora que nós, humanos, nos extinguiríamos, talvez devido a uma catástrofe natural ou a algo que provocámos. Outras espécies sobreviveriam, e dentro de

cinquenta milhões de anos, uma delas tornar-se-ia inteligente. Essa espécie quererá certamente saber tudo o que puder sobre a época dos humanos há muito desaparecida. Estaria especialmente interessada em conhecer a extensão do nosso conhecimento — e o que nos aconteceu.

Se a espécie humana se extinguir, então, dentro de apenas um milhão de anos, quase todos os registos detalhados da nossa civilização provavelmente desaparecerão. Permanecerão enterrados vestígios de algumas das nossas cidades e grandes infraestruturas, mas praticamente todos os documentos, filmes e gravações deixarão de existir. Futuros arqueólogos não humanos terão dificuldade em reconstruir a nossa história, tal como os paleontólogos de hoje se esforçam por compreender o que aconteceu aos dinossauros.

Como parte do nosso plano de legado, poderíamos preservar o nosso conhecimento de forma mais permanente — numa forma que perdure por dezenas de milhões de anos. Há várias maneiras de o fazer. Por exemplo, poderíamos arquivar continuamente uma base de dados de conhecimento como a Wikipédia. A Wikipédia está em constante atualização, o que permitiria documentar os acontecimentos até ao momento em que a nossa sociedade começasse a colapsar. Abrange uma vasta gama de temas, e o processo de arquivo poderia ser automatizado. Esse arquivo não deveria ficar localizado na Terra, já que esta poderia ser parcialmente destruída por um evento singular — e, ao longo de milhões de anos, muito pouco permaneceria intacto. Para resolver esse problema, poderíamos colocar o nosso arquivo numa série de satélites em órbita do Sol. Desta forma, o arquivo seria fácil de descobrir, mas difícil de alterar ou destruir fisicamente.

Projetaríamos o arquivo baseado em satélites de forma a permitir atualizações automáticas, mas sem que o conteúdo pudesse ser apagado. Os componentes eletrónicos desses satélites deixariam de funcionar pouco depois da nossa extinção, por isso, para ler o arquivo, uma futura espécie inteligente teria de desenvolver a tecnologia para viajar até ele, trazê-lo de volta à Terra e extrair os dados. Poderíamos usar vários satélites em órbitas distintas para garantir redundância. Já possuímos a

capacidade de criar e recuperar um arquivo deste tipo. Imagine se uma espécie inteligente anterior à nossa tivesse colocado satélites à volta do sistema solar — já os teríamos descoberto e trazido de volta à Terra.

Essencialmente, poderíamos criar uma cápsula do tempo desenhada para durar milhões ou centenas de milhões de anos. Num futuro longínquo, seres inteligentes — quer evoluam na Terra ou venham de outra estrela — poderiam descobrir essa cápsula do tempo e ler o seu conteúdo. Nunca saberemos se o nosso repositório será descoberto ou não — é essa a natureza dos planos de legado. Mas, se o for, e for lido no futuro, imagine o quão gratos seriam os seus recetores. Basta pensarmos no entusiasmo que sentiríamos se descobríssemos algo semelhante.

Um plano de legado para a humanidade é semelhante ao plano de legado de um indivíduo. Gostaríamos que a nossa espécie durasse para sempre — e talvez isso aconteça. Mas é sensato preparar um plano, caso o milagre não se concretize. Já sugeri várias ideias que poderíamos levar a cabo. Uma é arquivar a nossa história e conhecimento de forma que futuras espécies inteligentes na Terra possam aprender sobre a humanidade — o que sabíamos, a nossa história e o que nos aconteceu, afinal. Outra é criar um sinal duradouro, que diga a seres inteligentes noutros pontos do espaço e do tempo que existiu, outrora, vida inteligente em torno da estrela que chamamos Sol. A beleza de um sinal duradouro é que poderá, inclusive, ajudar-nos no curto prazo — ao levar-nos a descobrir que outras espécies inteligentes nos antecederam.

Vale a pena investir tempo e dinheiro em iniciativas como esta? Não seria melhor canalizar todos os nossos esforços para melhorar a vida na Terra? Há sempre um atrito entre investir no curto prazo e investir no longo prazo. Os problemas imediatos são mais urgentes, enquanto que os investimentos no futuro oferecem poucos benefícios imediatos. Todas as organizações — seja um governo, uma empresa ou uma família — enfrentam este dilema. Contudo, não investir no longo prazo é garantir o fracasso futuro. Neste caso, acredito que investir num plano de legado para a humanidade traz vários benefícios a curto prazo. Manter-nos-á mais conscientes das ameaças existenciais que enfrentamos. Levará

mais pessoas a pensar nas consequências a longo prazo das nossas ações como espécie. E oferecerá um tipo de propósito à nossa existência — caso, eventualmente, falhemos.

CAPÍTULO 16

Genes versus Conhecimento

"Cérebro Antigo—Cérebro Novo" é o título do primeiro capítulo deste livro. É também um tema subjacente. Recordemos que 30 por cento do nosso cérebro, o cérebro antigo, é composto por muitas partes distintas. Estas áreas do cérebro antigo controlam as nossas funções corporais, comportamentos básicos e emoções. Alguns desses comportamentos e emoções levam-nos a ser agressivos, violentos, cobiçosos, a mentir e a enganar. Todos nós albergamos estas tendências, em maior ou menor grau, porque a evolução descobriu que são úteis para propagar genes. Setenta por cento do nosso cérebro, o cérebro novo, é composto por uma única coisa: o neocórtex. O neocórtex aprende um modelo do mundo, e é esse modelo que nos torna inteligentes. A inteligência evoluiu porque, tal como os comportamentos instintivos, também é útil para propagar genes. Estamos aqui ao serviço dos nossos genes, mas o equilíbrio de poder entre o cérebro antigo e o cérebro novo começou a mudar.

Durante milhões de anos, os nossos antepassados tiveram um conhecimento limitado do nosso planeta e do universo mais vasto. Compreendiam apenas aquilo que podiam experienciar diretamente. Não sabiam qual o tamanho da Terra, nem que esta era uma esfera. Não sabiam o que eram o Sol, a Lua, os planetas e as estrelas, nem por que razão se moviam no céu como o fazem. Não faziam ideia de quão antiga é a Terra nem de como surgiram as suas formas de vida. Os nossos antepassados desconheciam os factos mais básicos da nossa existência. Criavam histórias para explicar estes mistérios, mas essas histórias não eram verdadeiras.

Recentemente, utilizando a nossa inteligência, não só resolvemos os enigmas que atormentavam os nossos antepassados, como o ritmo das descobertas científicas tem vindo a acelerar. Sabemos quão incrivelmente vasto é o universo e quão

incrivelmente pequenos somos nós. Compreendemos agora que o nosso planeta tem milhares de milhões de anos e que a vida na Terra tem vindo também a evoluir há milhares de milhões de anos. Felizmente, tudo indica que o universo inteiro funciona segundo um conjunto único de leis — algumas das quais já descobrimos. Parece até possível, de forma sedutora, que possamos vir a descobrir todas essas leis. Milhões de pessoas em todo o mundo trabalham ativamente na investigação científica em geral, e milhares de milhões sentem-se ligadas a essa missão. É, verdadeiramente, uma época extraordinária para se estar vivo.

Contudo, enfrentamos um problema que poderá travar rapidamente esta nossa corrida para a iluminação — e até levar à extinção da nossa espécie. Como referi anteriormente, por mais inteligentes que nos tornemos, o nosso neocórtex permanece ligado ao cérebro antigo. À medida que as nossas tecnologias se tornam cada vez mais poderosas, os comportamentos egoístas e de visão curta do cérebro antigo podem conduzir-nos à extinção ou mergulhar-nos no colapso social e numa nova idade das trevas. A agravar este risco está o facto de milhares de milhões de humanos ainda acreditarem em falsidades sobre os aspetos mais fundamentais da vida e do universo. Crenças falsas virais constituem outra fonte de comportamentos que ameaçam a nossa sobrevivência.

Encontramo-nos perante um dilema. "Nós" — o modelo inteligente de nós próprios que habita no neocórtex — estamos presos. Presos num corpo que não só está programado para morrer, como se encontra, em grande medida, sob o controlo de um bruto ignorante: o cérebro antigo. Podemos usar a nossa inteligência para imaginar um futuro melhor e tomar medidas para concretizar esse futuro desejado. Mas o cérebro antigo pode arruinar tudo. Ele gera comportamentos que, no passado, ajudaram os genes a replicar-se, mas muitos desses comportamentos não são nada nobres. Procuramos controlar os impulsos destrutivos e divisionistas do nosso cérebro antigo, mas, até agora, não fomos capazes de o fazer de forma plena. Muitos países na Terra continuam a ser governados por autocratas e ditadores cujas motivações são, em grande parte, ditadas pelo cérebro antigo: riqueza, sexo e domínio do tipo "macho alfa". Os movimentos populistas que sustentam esses autocratas

baseiam-se também em traços do cérebro antigo, como o racismo e a xenofobia.

Que fazer perante isto? No capítulo anterior, explorei formas de preservar o nosso conhecimento, caso a humanidade não sobreviva. Neste capítulo final, apresento três métodos que poderíamos seguir para evitar a nossa queda. O primeiro poderá ou não funcionar sem modificarmos os nossos genes; o segundo baseia-se na modificação genética; e o terceiro abandona por completo a biologia.

Estas ideias poderão parecer-lhe extremas. No entanto, pergunte a si próprio: qual é o propósito de viver? O que estamos a tentar preservar quando lutamos pela sobrevivência? No passado, viver foi sempre sobre preservar e replicar genes — quer o soubéssemos, quer não. Mas será esse o melhor caminho a seguir? E se decidirmos, em vez disso, que viver deve centrar-se na inteligência e na preservação do conhecimento? Se fizermos essa escolha, aquilo que hoje consideramos extremo poderá ser, no futuro, simplesmente a decisão mais lógica.

As três ideias que aqui apresento são, a meu ver, possíveis e têm uma forte probabilidade de virem a ser seguidas. Podem parecer improváveis agora — tal como os computadores portáteis pareciam improváveis em 1992. Teremos de deixar o tempo seguir o seu curso para vermos qual delas, se alguma, se revelará viável.

Tornar-se uma Espécie Multiplanetária

Quando o nosso Sol morrer, toda a vida no nosso sistema solar morrerá também. No entanto, a maioria dos eventos de extinção que nos preocupam seriam localizados na Terra. Se, por exemplo, um grande asteroide atingisse a Terra, ou se a tornássemos inabitável através de uma guerra nuclear, os outros planetas próximos não seriam afetados. Assim, uma forma de reduzir o risco de extinção é tornarmo-nos uma espécie de dois planetas. Se conseguíssemos estabelecer uma presença permanente noutro planeta ou lua próximos, então a nossa espécie e o conhecimento que acumulámos poderiam sobreviver mesmo que a Terra se tornasse inabitável. Esta lógica é uma das forças motrizes por

detrás dos esforços atuais para levar pessoas a Marte, que parece ser a melhor opção para o estabelecimento de uma colónia humana. A possibilidade de viajar para outros planetas entusiasma-me. Já passou muito tempo desde que viajámos para destinos verdadeiramente novos e inexplorados.

A principal dificuldade de viver em Marte é que Marte é um lugar terrível para se viver. A ausência de uma atmosfera significativa significa que uma breve exposição ao exterior o matará, e uma fuga no teto ou uma janela partida poderá matar toda a sua família. A radiação solar é mais intensa em Marte e constitui também um risco sério à vida humana, o que implicaria uma proteção constante contra o Sol. O solo marciano é venenoso e não há água à superfície. Sinceramente, é mais fácil viver no Pólo Sul do que em Marte. Mas isso não significa que devamos desistir da ideia. Acredito que poderíamos viver em Marte, mas para isso precisamos de algo que ainda não temos: robôs inteligentes e autónomos.

Para os seres humanos viverem em Marte, seriam necessários edifícios grandes e herméticos, onde se pudesse habitar e cultivar alimentos. Precisaríamos de extrair água e minerais de minas e de fabricar ar respirável. Em última instância, seria necessário terraformar Marte para reintroduzir uma atmosfera. Estes são projetos de infraestrutura gigantescos, que poderiam demorar décadas ou séculos a concretizar. Até que Marte se tornasse autossuficiente, tudo o que fosse necessário teria de ser enviado: comida, ar, água, medicamentos, ferramentas, equipamento de construção, materiais e pessoas — muitas pessoas. Todo o trabalho teria de ser feito com fatos espaciais volumosos. É difícil exagerar as dificuldades que os humanos enfrentariam ao tentar construir ambientes habitáveis e toda a infraestrutura necessária para criar uma colónia marciana permanente e autossustentável. A perda de vidas, os danos psicológicos e os custos financeiros seriam imensos — provavelmente superiores ao que estamos dispostos a suportar.

No entanto, preparar Marte para os humanos poderia ser viável se, em vez de enviarmos engenheiros e operários humanos, enviássemos engenheiros e operários robóticos inteligentes. Esses robôs obteriam energia do Sol e poderiam trabalhar no exterior sem

precisar de comida, água ou oxigénio. Poderiam laborar incansavelmente, sem stress emocional, durante o tempo necessário para tornar Marte habitável para os humanos. Esta força de trabalho robótica teria de operar maioritariamente de forma autónoma. Se dependessem de uma comunicação constante com a Terra, o progresso seria demasiado lento.

Nunca fui um apreciador da literatura de ficção científica, e este cenário soa suspeitosamente a ficção. No entanto, não vejo razão para que não o possamos concretizar e, se quisermos tornar-nos uma espécie multiplanetária, acredito que não temos escolha. Para que os humanos possam viver em Marte de forma permanente, é necessário o auxílio de máquinas inteligentes. O requisito fundamental é dotar a força de trabalho robótica em Marte do equivalente a um neocórtex. Os robôs precisam de utilizar ferramentas complexas, manipular materiais, resolver problemas imprevistos e comunicar entre si, de forma semelhante aos humanos. Acredito que a única maneira de alcançar isto é completar a engenharia reversa do neocórtex e criar estruturas equivalentes em silício. Os robôs autónomos precisam de ter um cérebro construído com base nos princípios que descrevi anteriormente — os princípios da Teoria dos Mil Cérebros da Inteligência.

Criar robôs verdadeiramente inteligentes é algo ao nosso alcance, e estou certo de que acontecerá. Creio que o poderíamos conseguir em poucas décadas, se fizéssemos disso uma prioridade. Felizmente, existem também muitas razões terrestres para desenvolver robôs inteligentes. Por isso, mesmo que não o tornemos uma prioridade nacional ou internacional, as forças de mercado acabarão por financiar o desenvolvimento da inteligência artificial e da robótica. Espero que as pessoas em todo o mundo venham a compreender que tornar-se uma espécie multiplanetária é um objetivo entusiasmante e importante para a nossa sobrevivência — e que trabalhadores robóticos inteligentes são essenciais para o alcançar.

Mesmo que criemos trabalhadores robóticos inteligentes, terraformemos Marte e estabeleçamos colónias humanas, continuaremos a ter um problema. Os humanos que forem para Marte serão iguais aos humanos que habitam a Terra. Terão um

cérebro antigo e todas as complicações e riscos que isso implica. Os humanos em Marte lutarão por território, tomarão decisões baseadas em crenças falsas e provavelmente criarão novos riscos existenciais para os que lá habitarem.

A história sugere que, eventualmente, os habitantes de Marte e os da Terra acabarão por entrar em conflito de formas que poderão pôr em perigo uma ou ambas as populações. Por exemplo, imagine que, daqui a duzentos anos, dez milhões de humanos vivem em Marte. Mas depois, algo corre mal na Terra. Talvez contaminemos acidentalmente a maior parte do planeta com elementos radioativos, ou o clima terrestre entre em colapso abruptamente. O que aconteceria? Milhares de milhões de habitantes da Terra poderiam, subitamente, querer mudar-se para Marte. Se deixarmos a imaginação fluir um pouco, vemos facilmente como isso poderia acabar mal para todos. Não pretendo especular sobre desfechos negativos, mas é importante reconhecer que tornar-se uma espécie multiplanetária não é uma solução milagrosa. Os humanos são humanos, e os problemas que criamos na Terra existirão também noutros planetas que venhamos a habitar.

E quanto a tornar-se uma espécie multiestelar? Se os humanos pudessem colonizar outros sistemas estelares, então poderíamos expandir-nos por toda a galáxia, e a probabilidade de alguns dos nossos descendentes sobreviverem indefinidamente aumentaria drasticamente.

Será possível a viagem interestelar humana? Por um lado, parece que sim. Existem quatro estrelas a menos de cinco anos-luz de nós e onze estrelas a menos de dez anos-luz. Einstein demonstrou que é impossível acelerar até à velocidade da luz, por isso, admitamos que viajamos a metade dessa velocidade. Uma missão a uma estrela próxima poderia ser realizada em uma ou duas décadas. Por outro lado, não sabemos como atingir sequer algo próximo dessa velocidade. Com as tecnologias de que dispomos hoje, levaríamos dezenas de milhares de anos para alcançar a estrela mais próxima. Os humanos não podem suportar uma viagem tão longa.

Há muitos físicos a pensar em formas engenhosas de ultrapassar os problemas da viagem interestelar. Talvez venham a descobrir formas de viajar perto da velocidade da luz — ou até mais depressa. Muitas coisas que pareciam impossíveis há apenas duzentos anos são hoje banais. Imagine que, em 1820, discursava num encontro de cientistas e dizia que, no futuro, qualquer pessoa poderia viajar confortavelmente de continente para continente em poucas horas, ou que as pessoas manteriam conversas presenciais com outras pessoas, em qualquer parte do mundo, apenas olhando para a sua mão e falando para ela. Ninguém acreditaria que tais coisas seriam possíveis — mas cá estamos. O futuro surpreender-nos-á certamente com avanços novos, hoje inconcebíveis — e um desses avanços poderá ser a viagem espacial prática.

Ainda assim, sinto-me à vontade para prever que a viagem interestelar humana não acontecerá nos próximos cinquenta anos. E não me surpreenderia se nunca viesse a acontecer.

Continuo, no entanto, a defender que nos tornemos uma espécie multiplanetária. Será uma aventura exploratória inspiradora e poderá reduzir o risco, a curto prazo, de extinção humana. Mas os riscos e limitações inerentes à nossa herança evolutiva permanecem. Mesmo que consigamos estabelecer colónias em Marte, talvez tenhamos de aceitar que nunca iremos além do nosso sistema solar.

Contudo, temos outras opções. Essas exigem que nos observemos com objetividade e perguntemos: o que é que queremos realmente preservar da humanidade? Irei abordar essa questão em primeiro lugar, antes de apresentar mais duas opções para garantir o nosso futuro.

Escolher o Nosso Futuro

A partir do Iluminismo, no final do século XVIII, temos vindo a acumular evidências crescentes de que o universo progride sem uma mão orientadora. O surgimento da vida simples, seguido de organismos complexos e, por fim, da inteligência, não foi planeado nem inevitável. Do mesmo modo, o futuro da vida na Terra e o futuro da inteligência não estão pré-determinados. Tudo indica que

a única entidade no universo que se importa com a forma como o futuro se desenrola somos nós próprios. O único futuro desejável é aquele que desejamos.

Poderá contestar esta afirmação. Poderá dizer que existem muitas outras espécies a viver na Terra, algumas também dotadas de inteligência. Prejudicámos muitas dessas espécies e levámos outras à extinção. Não deveríamos considerar o que essas espécies "desejam"? Sim — mas a questão não é assim tão simples.

A Terra é dinâmica. As placas tectónicas que compõem a sua superfície estão em constante movimento, criando novas montanhas, novos continentes e novos mares, ao mesmo tempo que submergem estruturas existentes no centro do planeta. A vida é igualmente dinâmica. As espécies estão sempre a mudar. Não somos geneticamente idênticos aos nossos antepassados de há cem mil anos. A taxa de mudança pode ser lenta, mas nunca cessa. Se olharmos para a Terra sob esta perspetiva, então preservar espécies ou preservar a Terra deixa de fazer sentido. Não conseguimos travar as forças geológicas mais básicas da Terra, e também não conseguimos impedir que as espécies evoluam e se extingam.

Uma das minhas atividades preferidas é caminhar na natureza selvagem, e considero-me ambientalista. Mas não finjo que o ambientalismo se trata de preservar a natureza. Todo o ambientalista ficaria satisfeito com a extinção de certos seres vivos — por exemplo, o vírus da poliomielite — enquanto, ao mesmo tempo, faria grandes esforços para salvar uma flor selvagem em risco de extinção. Do ponto de vista do universo, esta é uma distinção arbitrária; nem o poliovírus nem a flor silvestre são melhores ou piores um do outro. A escolha do que proteger baseia-se no que é do nosso melhor interesse. O ambientalismo não é sobre preservar a natureza, mas sobre as escolhas que fazemos. Regra geral, os ambientalistas fazem escolhas que beneficiam os humanos do futuro. Procuramos abrandar as mudanças naquilo que apreciamos — como as zonas selvagens — para aumentar a probabilidade de os nossos descendentes também poderem desfrutar dessas coisas. Há outras pessoas que prefeririam transformar essas zonas em minas a céu aberto, para colher

benefícios imediatos — uma escolha mais típica do cérebro antigo. O universo é indiferente à opção que escolhermos. Cabe-nos a nós decidir se ajudamos os humanos do futuro ou os humanos do presente.

Não existe a opção de nada fazer. Como seres inteligentes, somos obrigados a fazer escolhas — e essas escolhas moldarão o futuro, para um lado ou para o outro. Quanto aos outros animais na Terra, podemos escolher ajudá-los ou não. Mas, enquanto aqui estivermos, não existe a possibilidade de deixar que as coisas sigam o seu curso "natural". Somos parte da natureza, e as nossas decisões terão impacto no futuro.

Na minha perspetiva, temos diante de nós uma escolha profunda. Trata-se de optar entre favorecer o cérebro antigo ou o cérebro novo. Mais especificamente: queremos que o nosso futuro seja conduzido pelos processos que nos trouxeram até aqui — nomeadamente, a seleção natural, a competição e o impulso dos genes egoístas? Ou queremos que o nosso futuro seja orientado pela inteligência e pelo desejo de compreender o mundo? Temos a oportunidade de escolher entre um futuro onde o motor principal é a criação e disseminação de conhecimento, e um futuro onde o motor principal é a cópia e disseminação de genes.

Para podermos exercer essa escolha, precisamos da capacidade de alterar o rumo da evolução através da manipulação genética, e da capacidade de criar inteligência sob forma não biológica. Já possuímos a primeira, e a segunda está iminente. O uso destas tecnologias tem suscitado debates éticos. Devemos manipular os genes de outras espécies para melhorar o nosso fornecimento alimentar? Devemos manipular os nossos próprios genes para "melhorar" a nossa descendência? Devemos criar máquinas inteligentes que sejam mais inteligentes e capazes do que nós?

Talvez já tenha opiniões formadas sobre estas questões. Pode achar que estas práticas são aceitáveis, ou pode considerá-las eticamente condenáveis. Seja como for, não vejo qualquer mal em discutirmos as nossas opções. Analisar cuidadosamente as escolhas que temos ajudará a tomar decisões informadas — independentemente do que decidirmos fazer.

Tornar-se uma espécie multiplanetária é uma tentativa de evitar a nossa extinção, mas ainda assim é um futuro ditado pelos genes. Que tipo de escolhas poderíamos fazer para favorecer a propagação do conhecimento em vez da propagação dos genes?

Modificar os Nossos Genes

Recentemente, desenvolvemos a tecnologia para editar moléculas de ADN com precisão. Em breve, ser-nos-á possível criar novos genomas e modificar os já existentes com a mesma facilidade e precisão com que se cria e edita um documento de texto. Os benefícios da edição genética podem ser imensos. Por exemplo, poderíamos eliminar doenças hereditárias que causam sofrimento a milhões de pessoas. Contudo, essa mesma tecnologia pode também ser utilizada para criar formas de vida inteiramente novas ou para modificar o ADN dos nossos filhos — por exemplo, tornando-os melhores atletas ou mais atraentes. Se consideramos este tipo de manipulação aceitável ou repugnante dependerá das circunstâncias. Modificar o nosso ADN para nos tornarmos mais atraentes pode parecer desnecessário, mas, se a edição genética for o único meio de impedir a extinção da nossa espécie, então torna-se um imperativo.

Imaginemos, por exemplo, que decidimos que estabelecer uma colónia em Marte é uma boa apólice de seguro para a sobrevivência a longo prazo da nossa espécie — e que muitas pessoas se oferecem para ir. Mas, em seguida, descobrimos que os seres humanos não conseguem viver em Marte por longos períodos devido à sua baixa gravidade. Já sabemos que passar meses em gravidade zero na Estação Espacial Internacional causa problemas de saúde. Talvez, após dez anos a viver sob a fraca gravidade marciana, os nossos corpos comecem a colapsar e morram. Uma população permanente em Marte pareceria então impossível. No entanto, imaginemos que conseguimos resolver este problema editando o genoma humano — e que as pessoas com essas modificações genéticas poderiam viver indefinidamente em Marte. Devemos permitir que as pessoas modifiquem os seus genes e os genes dos seus filhos para poderem viver em Marte? Quem quer que esteja disposto a mudar-se para Marte já está a aceitar riscos que põem a vida em causa. E os genes

das pessoas que viverem em Marte irão, de qualquer modo, mudar lentamente com o tempo. Então por que razão não hão de poder fazer essa escolha? Se acredita que este tipo de edição genética deve ser proibido, mudaria de opinião se a Terra se estivesse a tornar inabitável e a única forma de sobreviver fosse mudar-se para Marte?

Agora imagine que aprendemos a modificar os nossos genes para eliminar comportamentos agressivos e tornar uma pessoa mais altruísta. Devemos permitir isto? Pense que, quando selecionamos quem pode tornar-se astronauta, escolhemos pessoas que já possuem naturalmente esses atributos. E há boas razões para isso: aumenta a probabilidade de sucesso de uma missão espacial. Se, no futuro, enviarmos pessoas para viverem em Marte, provavelmente faremos um tipo de triagem semelhante. Não daríamos preferência a pessoas emocionalmente estáveis em vez de pessoas com pavio curto e histórico de agressividade? Quando um único ato irrefletido ou violento pode matar uma comunidade inteira, os habitantes de Marte não exigiriam, porventura, que os novos colonos passassem por algum tipo de teste de estabilidade emocional? Se pudéssemos criar um melhor cidadão através da edição do ADN, é provável que os habitantes de Marte insistissem nisso.

Consideremos mais um cenário hipotético. Existem peixes que conseguem sobreviver congelados no gelo. E se conseguíssemos modificar o nosso ADN para que um humano pudesse ser congelado da mesma forma e depois descongelado no futuro? Imagino que muitas pessoas gostariam de congelar o seu corpo para serem acordadas dali a cem anos. Seria entusiasmante viver os últimos dez ou vinte anos de vida no futuro. Permitiríamos tal coisa? E se esta modificação permitisse aos humanos viajar para outras estrelas? Mesmo que a viagem demorasse milhares de anos, os viajantes espaciais poderiam ser congelados na partida e descongelados à chegada ao destino. Não faltariam voluntários para uma viagem assim. Existirão razões válidas para proibir as modificações genéticas que tornariam possível tal missão?

Consigo imaginar inúmeros cenários em que poderíamos decidir que é do nosso interesse pessoal modificar

significativamente o nosso ADN. Não há um certo ou errado absoluto; existem apenas escolhas que podemos fazer. Se alguém diz que nunca se deve permitir a edição genética por princípio, então, quer se aperceba disso ou não, está a escolher um futuro que serve os interesses dos genes existentes — ou, como acontece com frequência, os das falsas crenças virais. Ao adotar tal posição, está a eliminar opções que poderiam ser do maior interesse para a sobrevivência a longo prazo da humanidade e do conhecimento.

Não estou a defender que se edite o genoma humano sem supervisão ou reflexão. E nada do que descrevi envolve coerção. Ninguém deve ser forçado a fazer nenhuma destas coisas. Limito-me a apontar que a edição genética é possível — e, portanto, temos escolhas a fazer. Pessoalmente, não vejo por que razão o caminho da evolução não orientada deve ser preferido ao de um percurso escolhido por nós. Podemos estar gratos aos processos evolutivos que nos trouxeram até aqui. Mas agora que aqui estamos, temos a opção de usar a nossa inteligência para tomar as rédeas do futuro. A nossa sobrevivência como espécie, e a sobrevivência do conhecimento, poderão estar mais asseguradas se o fizermos.

Um futuro desenhado através da edição do nosso ADN continua, contudo, a ser um futuro biológico — e isso impõe limites ao que é possível. Por exemplo, não está claro até onde poderemos ir com a edição genética. Será possível modificar o nosso genoma de forma a permitir que os humanos do futuro viajem entre as estrelas? Será possível criar humanos do futuro que não se matem uns aos outros num posto avançado remoto? Ninguém sabe. Atualmente, não temos conhecimento suficiente sobre o ADN para prever o que é possível e o que não é. E não me surpreenderia se descobríssemos que algumas das coisas que gostaríamos de fazer são, em princípio, impossíveis.

Passo agora à nossa última opção. Talvez seja a forma mais segura de garantir a preservação do conhecimento e a sobrevivência da inteligência — mas poderá também ser a mais difícil.

Sair da Órbita de Darwin

A forma suprema de libertar a nossa inteligência do domínio do cérebro antigo e da nossa biologia é criar máquinas que sejam inteligentes como nós, mas que não dependam de nós. Seriam agentes inteligentes capazes de viajar para além do nosso sistema solar e de sobreviver mais tempo do que nós. Essas máquinas partilhariam o nosso conhecimento, mas não os nossos genes. Se os humanos regredissem culturalmente — por exemplo, entrando numa nova Idade das Trevas — ou se nos extinguíssemos, a nossa descendência maquinal inteligente poderia continuar sem nós.

Hesito em usar a palavra "máquina", pois pode evocar a imagem de algo como um computador em cima de uma secretária, ou um robô humanoide, ou uma personagem maléfica de um conto de ficção científica. Tal como referi anteriormente, não conseguimos prever como serão as máquinas inteligentes do futuro — da mesma forma que os primeiros engenheiros de computadores não podiam imaginar como seriam os computadores do futuro. Ninguém, nos anos 1940, imaginava que os computadores poderiam ser mais pequenos do que um grão de arroz, suficientemente pequenos para serem integrados em praticamente tudo. Nem poderiam imaginar computadores poderosos baseados na nuvem, acessíveis em todo o lado, mas sem localização física definida.

Do mesmo modo, não conseguimos imaginar como serão as máquinas inteligentes do futuro, nem de que materiais serão feitas — por isso, não tentemos adivinhar. Isso poderia limitar a nossa capacidade de imaginar o que é realmente possível. Em vez disso, vamos considerar as duas razões pelas quais poderíamos querer criar máquinas inteligentes capazes de viajar até às estrelas… sem nós.

Objetivo Número Um: Preservar o Conhecimento

No capítulo anterior, descrevi como poderíamos preservar o conhecimento num repositório em órbita do Sol, a que chamei Wiki Terra. O repositório que descrevi era estático — semelhante a uma biblioteca de livros impressos a flutuar no espaço. O nosso objetivo ao criá-lo seria preservar conhecimento, com a esperança de que,

no futuro, algum agente inteligente o descobrisse e conseguisse decifrar o seu conteúdo. Contudo, sem humanos a garantir a sua manutenção, o repositório acabaria por deteriorar-se. A Wiki Terra não se copia a si própria, não se repara — e, portanto, é temporária. Projetá-la-íamos para durar muito tempo, mas chegaria o dia, no futuro longínquo, em que deixaria de ser legível.

O neocórtex humano também é como uma biblioteca: contém conhecimento sobre o mundo. Mas, ao contrário da Wiki Terra, o neocórtex cria cópias do que sabe, transferindo o conhecimento para outros seres humanos. Este livro, por exemplo, é uma tentativa minha de transferir para outras pessoas — como você — algumas coisas que sei. Isso assegura a distribuição do conhecimento. A perda de uma pessoa não acarreta a perda definitiva do conhecimento. A forma mais segura de preservar conhecimento é continuar a fazer cópias.

Por isso, um dos objetivos de criar máquinas inteligentes seria replicar aquilo que os humanos já fazem: preservar conhecimento através da criação e disseminação de cópias. Quereríamos usar máquinas inteligentes para esse fim porque elas poderiam continuar a preservar o conhecimento muito depois de nós termos desaparecido — e poderiam distribuir esse conhecimento por locais onde não conseguimos ir, como outras estrelas. Ao contrário dos humanos, as máquinas inteligentes poderiam espalhar-se lentamente pela galáxia. Idealmente, poderiam partilhar conhecimento com seres inteligentes noutros pontos do universo. Imagine a emoção de descobrir um repositório de conhecimento e de história galática que tivesse viajado até ao nosso sistema solar.

No capítulo anterior sobre o "testamento da espécie", descrevi tanto a ideia da Wiki Terra como a ideia de criar um sinal duradouro a indicar que nós, uma espécie inteligente, existimos uma vez no nosso sistema solar. Juntos, esses dois sistemas poderiam, potencialmente, orientar outros seres inteligentes até ao nosso sistema e, depois, até à descoberta do nosso repositório de conhecimento. O que proponho neste capítulo é uma forma diferente de alcançar um resultado semelhante: em vez de atrair inteligência alienígena até ao repositório de conhecimento no nosso sistema solar, enviamos cópias do nosso conhecimento e da nossa

história por toda a galáxia. Em qualquer dos casos, algo inteligente teria de fazer a longa viagem pelo espaço.

Tudo se desgasta. À medida que as máquinas inteligentes viajassem pelo espaço, algumas seriam danificadas, perdidas ou destruídas inadvertidamente. Por isso, a nossa descendência maquinal inteligente teria de ser capaz de se autorreparar e, quando necessário, de criar cópias de si própria. Sei que isto assustará quem teme que as máquinas inteligentes venham a dominar o mundo. Como já referi, não acredito que esse receio se justifique, dado que a maioria das máquinas inteligentes não será capaz de se reproduzir. Mas neste cenário, a autorreplicação é um requisito. No entanto, essa dificuldade em replicar-se é precisamente uma das razões pelas quais este cenário poderá não ser viável. Imaginemos um pequeno grupo de máquinas inteligentes a viajar pelo espaço. Após milhares de anos, chegam a um novo sistema solar. Encontram planetas estéreis e um com vida primitiva, unicelular. É isso que teria sido encontrado por um visitante no nosso sistema solar há milhares de milhões de anos. Suponhamos agora que essas máquinas decidem que precisam de substituir dois dos seus membros e de criar algumas novas unidades inteligentes para enviar para outra estrela. Como o fariam? Se, por exemplo, tivessem sido construídas com chips de silício como os que usamos nos computadores, então teriam de construir fábricas de chips de silício e recriar toda a cadeia de fornecimento necessária? Isso poderia ser inviável. Talvez venhamos a aprender como criar máquinas inteligentes que consigam replicar-se a partir de elementos comuns — tal como a vida baseada em carbono na Terra.

Não sei como superar os inúmeros problemas práticos que a viagem interestelar apresenta. Mas, mais uma vez, acredito que não devemos fixar-nos nas manifestações físicas das máquinas inteligentes do futuro. Poderão existir formas de construir essas máquinas usando materiais e métodos de fabrico que ainda não inventámos. Por agora, é mais importante discutir objetivos e conceitos que nos ajudem a decidir se é algo que desejaríamos fazer — caso fosse possível. Se decidirmos que enviar máquinas inteligentes para explorar a galáxia e espalhar conhecimento é algo

que queremos realizar, então é possível que consigamos conceber formas de ultrapassar os obstáculos.

Objetivo Número Dois: Adquirir Novo Conhecimento

Se conseguíssemos criar máquinas inteligentes autossustentáveis que viajassem entre as estrelas, elas fariam novas descobertas. Descobririam, sem dúvida, novos tipos de planetas e estrelas, e fariam outras descobertas que nem conseguimos imaginar. Talvez encontrassem respostas para os grandes mistérios do universo — como a sua origem ou o seu destino. Essa é a natureza da exploração: não sabemos o que vamos aprender, mas sabemos que algo será aprendido. Se enviássemos seres humanos para explorar a galáxia, esperaríamos que fizessem descobertas. E, em muitos aspetos, as máquinas inteligentes terão uma capacidade de descoberta superior à dos humanos. O seu equivalente cerebral terá mais memória, funcionará com maior velocidade e disporá de sensores inéditos. Seriam cientistas melhores do que nós. Se máquinas inteligentes atravessassem a nossa galáxia, aumentariam continuamente o corpo de conhecimento sobre o universo.

Um Futuro com Propósito e Direção

Durante muito tempo, os seres humanos sonharam com a possibilidade de viajar entre as estrelas. Porquê?

Uma razão é a de estender e preservar os nossos genes. Esta ideia assenta na noção de que o destino de uma espécie é explorar continuamente novas terras e estabelecer colónias onde quer que possa. Fizemos isso repetidamente ao longo da nossa história — atravessando montanhas e oceanos para fundar novas sociedades. Isso serve os interesses dos nossos genes e, por conseguinte, estamos programados para explorar. A curiosidade é uma das funções do cérebro antigo. É difícil resistir ao impulso de explorar, mesmo quando seria mais seguro não o fazer. Se os humanos conseguissem viajar até às estrelas, isso seria apenas uma

extensão natural daquilo que sempre fizemos: espalhar os nossos genes pelo maior número possível de lugares.

A segunda razão — a que sugeri neste capítulo — é a de estender e preservar o nosso conhecimento. Esta linha de pensamento baseia-se na suposição de que é a inteligência, e não os nossos genes em particular, que torna a nossa espécie importante. Assim, deveríamos viajar até às estrelas para aprender mais e salvaguardar o nosso conhecimento para o futuro.

Mas será essa uma escolha melhor? O que há de errado em continuarmos como até agora? Podemos ignorar toda esta conversa sobre preservar o conhecimento ou criar máquinas inteligentes. A vida na Terra tem sido, até aqui, bastante boa. Se os humanos não puderem viajar para outras estrelas… e depois? Por que não simplesmente continuar o percurso, aproveitando a viagem enquanto dura?

É uma escolha razoável, e no fim de contas poderá ser a única que tenhamos. Mas quero apresentar argumentos a favor do conhecimento em detrimento dos genes. Existe uma diferença fundamental entre ambos — uma diferença que torna, na minha opinião, mais nobre e valioso o objetivo de preservar e disseminar o conhecimento do que o de preservar e disseminar os nossos genes.

Os genes são apenas moléculas que se replicam. À medida que evoluem, não seguem nenhuma direção específica, nem há um gene que seja intrinsecamente melhor do que outro — tal como uma molécula não é melhor do que outra. Alguns genes podem ser mais eficientes na replicação, mas à medida que os ambientes mudam, os genes mais eficazes também mudam. O mais importante é que não há direção global nessas mudanças. A vida baseada em genes não tem rumo nem objetivo. A vida pode manifestar-se sob a forma de vírus, bactéria unicelular ou árvore. Mas nada indica que uma forma de vida seja superior a outra, a não ser na sua capacidade de se replicar.

O conhecimento é diferente. O conhecimento tem direção e um objetivo final. Por exemplo, consideremos a gravidade. Num passado não muito distante, ninguém sabia por que razão as coisas caíam para baixo e não para cima. Newton criou a primeira teoria

bem-sucedida da gravidade. Propôs que se tratava de uma força universal e demonstrou que obedecia a um conjunto de leis simples que podiam ser expressas matematicamente. Após Newton, nunca mais voltámos ao ponto de não termos uma teoria da gravidade. A explicação de Einstein é melhor do que a de Newton — e nunca mais voltaremos à teoria de Newton. Isso não significa que Newton estivesse errado. As suas equações continuam a descrever com precisão a gravidade tal como a experienciamos no dia-a-dia. A teoria de Einstein incorpora a de Newton, mas descreve melhor a gravidade em condições invulgares. O conhecimento tem uma direção: pode passar de nenhum conhecimento, para Newton, e de Newton para Einstein — mas não pode regredir.

Além de direção, o conhecimento tem um objetivo final. Os primeiros exploradores humanos não sabiam qual era a dimensão da Terra. Por mais que viajassem, havia sempre mais por descobrir. Seria a Terra infinita? Teria uma orla onde, ao prosseguir, se cairia no vazio? Ninguém sabia. Mas havia um objetivo: supunha-se que existia uma resposta para a pergunta "Qual é o tamanho da Terra?". Acabámos por alcançar esse objetivo com uma resposta surpreendente: a Terra é uma esfera, e hoje sabemos qual a sua dimensão.

Estamos hoje diante de mistérios semelhantes: Qual é o tamanho do universo? Estende-se até ao infinito? Tem um limite? Enrola-se sobre si mesmo como a Terra? Existem vários universos? Há muitas outras questões que ainda não compreendemos: O que é o tempo? Como surgiu a vida? Quão comum é a vida inteligente? Responder a estas perguntas é um objetivo — e a história sugere que somos capazes de o alcançar.

Um futuro orientado pelos genes tem pouca ou nenhuma direção e apenas objetivos de curto prazo: manter-se saudável, ter filhos, desfrutar da vida. Um futuro desenhado em função do conhecimento tem direção e objetivos finais.

A boa notícia é que não temos de escolher apenas um desses futuros. É possível fazer ambas as coisas. Podemos continuar a viver na Terra, fazendo o nosso melhor para mantê-la habitável e protegendo-nos dos nossos piores impulsos. E podemos, em simultâneo, dedicar recursos à preservação do conhecimento e à

continuidade da inteligência para um tempo futuro em que já não estejamos aqui.

Escrevi a Parte 3 deste livro — os últimos cinco capítulos — para defender o conhecimento em detrimento dos genes. Convidei-o a olhar para os humanos com objetividade. Convidei-o a ver como tomamos más decisões e por que é que os nossos cérebros são vulneráveis a falsas crenças. Convidei-o a considerar o conhecimento e a inteligência como algo mais precioso do que os genes e a biologia — e, por isso, dignos de preservação para além do seu atual domicílio no cérebro biológico. Convidei-o a considerar a possibilidade de uma descendência baseada na inteligência e no conhecimento — e que esses descendentes possam ser tão dignos quanto os descendentes baseados em genes.

Quero frisar novamente que não estou a prescrever o que devemos fazer. O meu objetivo é estimular a reflexão, mostrar que algumas das coisas que tomamos por certezas éticas são, na verdade, escolhas — e trazer à superfície ideias que têm estado indevidamente afastadas do centro do debate.

Agora, quero regressar ao presente.

Considerações Finais

Tenho uma Visão que Nunca Deixa de me Fascinar. Tenho uma visão que nunca deixa de me entreter. Imagino o vasto universo, com as suas centenas de milhares de milhões de galáxias. Cada galáxia contém centenas de milhares de milhões de estrelas. Em torno de cada estrela, visualizo planetas com uma variedade ilimitada de formas. Imagino estes triliões de objetos de dimensões monstruosas a orbitarem-se lentamente uns aos outros no imenso vazio do espaço, durante milhares de milhões de anos. O que me maravilha é que a única coisa no universo que sabe disto — a única coisa que sabe que o universo existe — é o nosso cérebro. Se não existissem cérebros, nada saberia que algo existe. E isso levanta a pergunta que referi no início do livro: Se não há conhecimento de algo, podemos dizer que essa coisa existe de facto? O facto de o nosso cérebro desempenhar um papel tão único é fascinante. Claro que podem existir seres inteligentes noutros pontos do universo — mas isso torna ainda mais estimulante pensar sobre o assunto.

Pensar sobre o universo e sobre a singularidade da inteligência é uma das razões pelas quais quis estudar o cérebro. Mas existem muitas outras razões aqui mesmo na Terra. Por exemplo, compreender como o cérebro funciona tem implicações para a medicina e para a saúde mental. Resolver os mistérios do cérebro levará à verdadeira inteligência artificial, o que beneficiará todos os setores da sociedade — tal como os computadores o fizeram — e proporcionará melhores métodos para ensinar as nossas crianças. Mas, em última análise, tudo regressa à nossa inteligência singular. Somos a espécie mais inteligente. Se queremos compreender quem somos, então temos de compreender como o cérebro cria a inteligência. Inverter o processo de engenharia do cérebro e compreender a inteligência é, na minha opinião, a mais importante missão científica que a humanidade alguma vez empreenderá.

Quando comecei esta missão, tinha uma compreensão limitada do que fazia o neocórtex. Eu e outros neurocientistas tínhamos algumas noções sobre o cérebro aprender um modelo do mundo —

mas eram ideias vagas. Não sabíamos como seria tal modelo nem como os neurónios poderiam criá-lo. Estávamos submersos em dados experimentais e era difícil dar sentido a esses dados sem um enquadramento teórico.

Desde então, neurocientistas de todo o mundo fizeram progressos significativos. Este livro foca-se naquilo que a minha equipa descobriu. Muito do que aprendemos foi surpreendente — como a revelação de que o neocórtex não contém apenas um modelo do mundo, mas cerca de 150 000 sistemas de modelação sensório-motora. Ou a descoberta de que tudo o que o neocórtex faz se baseia em quadros de referência.

Na primeira parte deste livro, descrevi a nova teoria sobre como o neocórtex funciona e como aprende um modelo do mundo. Chamamos-lhe Teoria dos Mil Cérebros da Inteligência. Espero ter conseguido expor essas ideias com clareza e que tenha achado os meus argumentos convincentes. A certa altura, ponderei terminar o livro ali mesmo. Um enquadramento teórico para compreender o neocórtex já seria, só por si, ambicioso o suficiente para um único livro. No entanto, compreender o cérebro conduz, naturalmente, a outras questões de grande relevância — e por isso continuei.

Na Parte 2, argumentei que a IA atual não é inteligente. A verdadeira inteligência exige que as máquinas aprendam um modelo do mundo da mesma forma que o neocórtex o faz. E defendi por que razão a inteligência artificial não representa um risco existencial, ao contrário do que muitos pensam. A inteligência artificial será uma das tecnologias mais benéficas que alguma vez criaremos. Tal como qualquer outra tecnologia, haverá quem a use de forma abusiva — e isso preocupa-me mais do que a IA em si. Por si só, a inteligência artificial não constitui uma ameaça existencial — e acredito que os benefícios superarão largamente os riscos.

Por fim, na Parte 3 do livro, examinei a condição humana através da lente da inteligência e da teoria do cérebro. Como provavelmente já percebeu, estou preocupado com o futuro. Preocupo-me com o bem-estar da sociedade humana e até com a sobrevivência a longo prazo da nossa espécie. Um dos meus objetivos é alertar para o facto de a combinação entre o cérebro

antigo e as falsas crenças representar um risco existencial real — muito mais sério do que o suposto perigo da IA. Abordei diferentes formas de reduzir os riscos que enfrentamos. Várias dessas estratégias exigem a criação de máquinas inteligentes.

Escrevi este livro para partilhar aquilo que eu e os meus colegas aprendemos sobre a inteligência e o cérebro. Mas para além de transmitir esta informação, espero conseguir inspirar alguns de vocês a agir com base nela. Se é jovem, ou se está a ponderar uma mudança de carreira, considere entrar nos campos da neurociência e da inteligência artificial. Poucos assuntos são tão interessantes, tão desafiantes e tão importantes. Contudo, devo avisá-lo: será difícil se quiser seguir as ideias que descrevi neste livro. Tanto a neurociência como a aprendizagem automática são áreas vastas e com grande inércia institucional. Não tenho dúvidas de que os princípios que aqui expus virão a desempenhar papéis centrais em ambas — mas pode demorar anos até isso acontecer. Até lá, será preciso determinação e engenho.

Tenho ainda mais um pedido — este, para todos. Espero que um dia cada pessoa na Terra aprenda como funciona o seu cérebro. Para mim, isto deveria ser algo natural, como: "Ah, tens um cérebro? Então eis o que precisas de saber sobre ele." A lista de coisas que todos deveriam saber é curta. Eu incluiria o facto de o cérebro ser composto por uma parte nova e por partes mais antigas. Incluiria o modo como o neocórtex aprende um modelo do mundo, enquanto as partes mais antigas geram as nossas emoções e comportamentos mais primitivos. Incluiria o facto de o cérebro antigo poder assumir o controlo, levando-nos a agir de formas que sabemos serem erradas. E incluiria o facto de todos nós sermos suscetíveis a falsas crenças — e que algumas crenças são virais.

Acredito que todos deveriam saber estas coisas — da mesma forma que todos deveriam saber que a Terra orbita o Sol, que as moléculas de ADN codificam os nossos genes, e que os dinossauros viveram na Terra durante milhões de anos e estão agora extintos. Isto é importante. Muitos dos problemas que enfrentamos — desde guerras até às alterações climáticas — são provocados por falsas crenças ou pelos desejos egoístas do cérebro antigo, ou por ambos. Se cada ser humano compreendesse o que se passa dentro da sua

cabeça, acredito que haveria menos conflitos — e um futuro mais luminoso à nossa frente.

Cada um de nós pode contribuir para esse esforço. Se é pai ou mãe, ensine os seus filhos sobre o cérebro, tal como o faria ao levantar uma laranja e uma maçã para lhes explicar o sistema solar. Se escreve livros infantis, considere escrever sobre o cérebro e as crenças. Se é educador, pergunte-se de que forma a teoria do cérebro pode ser integrada no currículo base. Muitas escolas já ensinam genética e tecnologias do ADN como parte do ensino secundário. Acredito que a teoria do cérebro é, no mínimo, igualmente importante — se não mais.

O que Somos Nós?

Como viemos aqui parar?

Qual é o nosso destino?

Durante milénios, os nossos antepassados fizeram estas perguntas fundamentais. É natural. Acordamos e damos por nós num mundo complexo e misterioso. Não há manual de instruções para a vida, nem uma história de fundo que nos explique do que se trata tudo isto. Fazemos o melhor que podemos para dar sentido à nossa situação — mas, durante a maior parte da história humana, fomos ignorantes. A partir de há algumas centenas de anos, começámos a responder a algumas destas questões fundamentais. Hoje compreendemos a química subjacente a todos os seres vivos. Compreendemos os processos evolutivos que conduziram à nossa espécie. E sabemos que a nossa espécie continuará a evoluir — e que provavelmente se extinguirá algures no futuro.

Perguntas semelhantes podem ser feitas sobre nós, enquanto seres mentais:

O que nos torna inteligentes e autoconscientes?

Como se tornou a nossa espécie inteligente?

Qual é o destino da inteligência e do conhecimento?

Espero tê-lo convencido de que estas perguntas não só têm resposta, como estamos a fazer excelentes progressos em encontrá-la. Espero também tê-lo convencido de que devemos preocupar-nos com o futuro da inteligência e do conhecimento — independentemente da nossa preocupação com o futuro da espécie humana. A nossa inteligência superior é única — e, tanto quanto sabemos, o cérebro humano é a única coisa no universo que sabe que o próprio universo existe. É a única coisa que conhece a dimensão do universo, a sua idade e as leis pelas quais ele se rege. Isto torna a nossa inteligência e o nosso conhecimento dignos de preservação. E dá-nos esperança de que, um dia, possamos compreender tudo.

Somos *Homo sapiens* — os humanos sábios. Esperemos ser suficientemente sábios para reconhecer o quão especiais somos. Suficientemente sábios para fazer as escolhas que assegurem a sobrevivência da nossa espécie durante o máximo de tempo possível aqui na Terra. E suficientemente sábios para fazer as escolhas que garantam que a inteligência e o conhecimento sobrevivam ainda por mais tempo — aqui na Terra, e por todo o universo.

Leituras Sugeridas

Pessoas que ouviram falar do nosso trabalho perguntam-me frequentemente o que recomendo ler para aprender mais sobre a Teoria dos Mil Cérebros e a neurociência relacionada. Esta pergunta costuma arrancar-me um suspiro profundo, porque não há uma resposta simples — e, para ser honesto, é difícil ler artigos de neurociência. Antes de lhe dar recomendações de leitura específicas, deixo-lhe algumas sugestões gerais:

A neurociência é um campo de estudo tão vasto que, mesmo sendo um cientista profundamente familiarizado com uma das suas subáreas, pode ter dificuldades em ler a literatura científica de outra. E se estiver completamente a começar, pode ser difícil saber por onde começar.

Se quiser aprender sobre um tema específico — digamos, colunas corticais ou células de grelha — e ainda não estiver familiarizado com esse tema, recomendo que comece com uma fonte como a Wikipédia. A Wikipédia costuma ter vários artigos sobre qualquer tópico, e pode navegar rapidamente entre eles através dos links. É a forma mais rápida que conheço para ganhar uma noção da terminologia, ideias, temas, etc. Vai notar que diferentes artigos às vezes discordam entre si ou usam terminologias diferentes. Encontrará essas mesmas divergências em artigos científicos revistos por pares. Como regra, é preciso ler várias fontes para se obter uma noção do que se sabe sobre um determinado tema.

Se quiser aprofundar mais, recomendo-lhe a leitura de artigos de revisão. Os artigos de revisão são publicados em revistas científicas com revisão por pares, mas como o nome indica, apresentam uma visão geral de um tema — incluindo as áreas onde os cientistas discordam. Costumam ser mais fáceis de ler do que os artigos típicos. Além disso, as referências bibliográficas são valiosas porque reúnem, numa só lista, a maioria dos artigos importantes sobre o tema. Uma boa forma de encontrar artigos de revisão é

usar um motor de busca como o Google Scholar e escrever algo como "review article for grid cells".

Só depois de aprender a nomenclatura, a história e os conceitos de um tema é que recomendo ler artigos científicos individuais. O título e o resumo de um artigo raramente são suficientes para saber se contém a informação que procura. O que costumo fazer é ler o resumo. Depois, examino as imagens, que, num bom artigo, devem contar a mesma história que o texto. Em seguida, salto para a secção de discussão no final. Esta secção é muitas vezes o único lugar onde os autores descrevem claramente o que o artigo realmente quer dizer. Só depois destes passos preliminares é que considero ler o artigo do princípio ao fim.

Abaixo apresento sugestões de leitura organizadas por temas. Existem centenas ou milhares de artigos sobre cada assunto, por isso só lhe posso dar algumas sugestões para começar.

Colunas Corticais

A Teoria dos Mil Cérebros baseia-se na proposta de Vernon Mountcastle de que as colunas corticais possuem arquiteturas semelhantes e desempenham funções semelhantes. A primeira referência abaixo é o ensaio original de Mountcastle, no qual ele propõe a ideia de um algoritmo cortical comum. A segunda é um artigo mais recente do mesmo autor, onde ele apresenta diversos resultados experimentais que apoiam a sua proposta. A terceira referência, de Buxhoeveden e Casanova, é uma revisão relativamente fácil de ler. Embora se foque sobretudo nas minicolunas, discute vários argumentos e evidências relacionados com a afirmação de Mountcastle. A quarta referência, de Thomson e Lamy, é um artigo de revisão sobre a anatomia cortical. É uma análise minuciosa das camadas celulares e das conexões prototípicas entre elas. É complexa, mas é um dos meus artigos favoritos.

Mountcastle, Vernon. "An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System." In The Mindful Brain, editado por Gerald M. Edelman e Vernon B. Mountcastle, 7–50. Cambridge, MA: MIT Press, 1978.

Mountcastle, Vernon. "The Columnar Organization of the Neocortex." Brain 120 (1997): 701–722.

Buxhoeveden, Daniel P., e Manuel F. Casanova. "The Minicolumn Hypothesis in Neuroscience." Brain 125, nº 5 (Maio 2002): 935–951.

Thomson, Alex M., e Christophe Lamy. "Functional Maps of Neocortical Local Circuitry." Frontiers in Neuroscience 1 (Outubro 2007): 19–42.

Hierarquia Cortical

O primeiro artigo listado abaixo, de Felleman e Van Essen, é aquele referido no Capítulo 1 e que descreveu pela primeira vez a hierarquia das regiões no neocórtex do macaco. Incluo-o sobretudo pelo seu valor histórico. Infelizmente, não está disponível em acesso aberto.

A segunda referência, de Hilgetag e Goulas, oferece uma visão mais atual das questões relacionadas com a hierarquia no neocórtex. Os autores enumeram diversos problemas em interpretar o neocórtex como uma hierarquia estrita.

A terceira referência, um artigo de Murray Sherman e Ray Guillery, defende que a principal via de comunicação entre duas regiões corticais é feita através de uma parte do cérebro chamada tálamo. A Figura 3 do artigo ilustra bem essa ideia. A proposta de Sherman e Guillery é muitas vezes ignorada por outros neurocientistas. Por exemplo, nenhuma das duas primeiras referências menciona as conexões através do tálamo. Embora eu não tenha abordado o tálamo neste livro, ele está tão intimamente ligado ao neocórtex que o considero uma extensão do próprio neocórtex. Os meus colegas e eu discutimos uma possível explicação para a via talâmica no nosso artigo "Frameworks", publicado em 2019, que será mencionado adiante.

Felleman, Daniel J., e David C. Van Essen. "Distributed Hierarchical Processing in the Primate Cerebral Cortex." Cerebral Cortex 1, nº 1 (Janeiro–Fevereiro 1991): 1.

Hilgetag, Claus C., e Alexandros Goulas. "'Hierarchy' in the Organization of Brain Networks." Philosophical Transactions of the Royal Society B: Biological Sciences 375, nº 1796 (Abril 2020).

Sherman, S. Murray, e R. W. Guillery. "Distinct Functions for Direct and Transthalamic Corticocortical Connections." Journal of Neurophysiology 106, nº 3 (Setembro 2011): 1068–1077.

Vias "O Quê" e "Onde"

No Capítulo 6, descrevi como as colunas corticais, baseadas em referenciais, podem ser aplicadas às vias "o quê" e "onde" no neocórtex. O primeiro artigo, de Ungerleider e Haxby, é um dos textos originais sobre este tema. O segundo artigo, de Goodale e Milner, apresenta uma descrição mais moderna. Nele, os autores defendem que uma terminologia mais adequada do que "o quê" e "onde" seria "perceção" e "ação". Este artigo não está disponível em acesso aberto. O terceiro artigo, de Rauschecker, é talvez o mais acessível para leitura.

Ungerleider, Leslie G., e James V. Haxby. "'What' and 'Where' in the Human Brain." Current Opinion in Neurobiology 4 (1994): 157–165.

Goodale, Melvyn A., e A. David Milner. "Two Visual Pathways—Where Have They Taken Us and Where Will They Lead in Future?" Cortex 98 (Janeiro 2018): 283–292.

Rauschecker, Josef P. "Where, When, and How: Are They All Sensorimotor? Towards a Unified View of the Dorsal Pathway in Vision and Audition." Cortex 98 (Janeiro 2018): 262–268.

Espigões Dendríticos

No Capítulo 4, abordei a nossa teoria de que os neurónios no neocórtex fazem previsões usando espigões dendríticos. Eis três artigos de revisão que exploram este tema. O primeiro, de London e Häusser, é talvez o mais acessível. O segundo, de Antic et al., é mais diretamente relevante para a nossa teoria, tal como o terceiro, de Major, Larkum e Schiller.

London, Michael, e Michael Häusser. "Dendritic Computation." Annual Review of Neuroscience 28, n.º 1 (Julho 2005): 503–532.

Antic, Srdjan D., Wen-Liang Zhou, Anna R. Moore, Shaina M. Short, e Katerina D. Ikonomu. "The Decade of the Dendritic NMDA Spike." Journal of Neuroscience Research 88 (Novembro 2010): 2991–3001.

Major, Guy, Matthew E. Larkum, e Jackie Schiller. "Active Properties of Neocortical Pyramidal Neuron Dendrites." Annual Review of Neuroscience 36 (Julho 2013): 1–24.

Células de Grade e Células de Lugar

Um elemento-chave da Teoria dos Mil Cérebro é que cada coluna cortical aprende modelos do mundo usando referenciais. Propomos que o neocórtex o faz com mecanismos semelhantes aos das células de grade e células de lugar no córtex entorrinal e no hipocampo. Para uma excelente visão geral sobre estas células, recomendo assistir ou ler as palestras Nobel de O'Keefe e dos Mosers, na ordem em que foram apresentadas. Os três coordenaram as suas intervenções para formar um conjunto articulado.

O'Keefe, John. "Spatial Cells in the Hippocampal Formation." Palestra Nobel. Gravada a 7 de Dezembro de 2014, na Aula Medica, Karolinska Institutet, Estocolmo. Vídeo, 45:17. www.nobelprize.org/prizes/medicine/2014/okeefe/lecture/

Moser, Edvard I. "Grid Cells and the Entorhinal Map of Space." Palestra Nobel. Gravada a 7 de Dezembro de 2014. Vídeo, 49:23. www.nobelprize.org/prizes/medicine/2014/edvard-moser/lecture/

Moser, May-Britt. "Grid Cells, Place Cells and Memory." Palestra Nobel. Gravada a 7 de Dezembro de 2014. Vídeo, 49:48. www.nobelprize.org/prizes/medicine/2014/may-britt-moser/lecture/

Células de Grade no Neocórtex

Apenas começámos a observar evidências de mecanismos de células de grade no neocórtex. No Capítulo 6, descrevi duas experiências de fMRI que mostraram sinais de células de grade em humanos durante tarefas cognitivas. Os dois primeiros artigos — de Doeller, Barry e Burgess, e de Constantinescu, O'Reilly e Behrens — descrevem essas experiências. O terceiro artigo, de Jacobs et al., relata resultados semelhantes em humanos durante cirurgias cerebrais abertas.

Doeller, Christian F., Caswell Barry, e Neil Burgess. "Evidence for Grid Cells in a Human Memory Network." Nature 463, n.º 7281 (Fevereiro 2010): 657–661.

Constantinescu, Alexandra O., Jill X. O'Reilly, e Timothy E. J. Behrens. "Organizing Conceptual Knowledge in Humans with a Gridlike Code." Science 352, n.º 6292 (Junho 2016): 1464–1468.

Jacobs, Joshua, et al. "Direct Recordings of Grid-Like Neuronal Activity in Human Spatial Navigation." Nature Neuroscience 16, n.º 9 (Setembro 2013): 1188–1190.

Artigos da Numenta sobre a Teoria dos Mil Cérebro

Este livro apresenta uma descrição geral da Teoria dos Mil Cérebro, mas não entra em muitos detalhes. Se quiser aprofundar, pode ler os artigos revistos por pares do meu laboratório. Contêm descrições pormenorizadas dos componentes, incluindo simulações e código-fonte. Todos os artigos são de acesso aberto. Artigo mais recente e o mais acessível:

Hawkins, Jeff, Marcus Lewis, Mirko Klukas, Scott Purdy, e Subutai Ahmad. "A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex." Frontiers in Neural Circuits 12 (Janeiro 2019): 121.

Artigo sobre espigões dendríticos como previsões e sinapses contextuais:

Hawkins, Jeff, e Subutai Ahmad. "Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex." Frontiers in Neural Circuits 10, n.º 23 (Março 2016): 1–13.

Primeira introdução da ideia de colunas que aprendem objetos inteiros:

Hawkins, Jeff, Subutai Ahmad, e Yuwei Cui. "A Theory of How Columns in the Neocortex Enable Learning the Structure of the World." Frontiers in Neural Circuits 11 (Outubro 2017): 81.

Extensão do artigo anterior, explicando em detalhe as representações de localização por células de grade:

Lewis, Marcus, Scott Purdy, Subutai Ahmad, e Jeff Hawkins. "Locations in the Neocortex: A Theory of Sensorimotor Object Recognition Using Cortical Grid Cells." Frontiers in Neural Circuits 13 (Abril 2019): 22.

Agradecimentos

Embora o meu nome figure como autor, este livro e a Teoria dos Mil Cérebro foram criados por muitas pessoas. Quero contar-vos quem são e os papéis que desempenharam.

A Teoria dos Mil Cérebro

Desde o início, mais de uma centena de colaboradores, pós-doutorandos, estagiários e cientistas visitantes trabalharam na Numenta. Todos contribuíram, de uma forma ou de outra, para a investigação e os artigos que produzimos. Se faz parte deste grupo, o meu agradecimento.

Há, no entanto, algumas pessoas que merecem menção especial. O Dr. Subutai Ahmad tem sido o meu parceiro científico há quinze anos. Para além de gerir a nossa equipa de investigação, contribui para as nossas teorias, cria simulações e desenvolve a maior parte da matemática subjacente ao nosso trabalho. Os avanços alcançados na Numenta não teriam ocorrido sem Subutai. Marcus Lewis também deu contributos importantes à teoria. Frequentemente assumia tarefas científicas difíceis e surgia com ideias surpreendentes e intuições profundas. Luiz Scheinkman é um engenheiro de software incrivelmente talentoso e foi um elemento-chave em tudo o que fizemos. Scott Purdy e o Dr. Yuwei Cui também deram contributos significativos à teoria e às simulações.

Teri Fry e eu trabalhámos juntos tanto no Redwood Neuroscience Institute como na Numenta. A Teri gere de forma exemplar o nosso escritório e tudo o que é necessário para manter uma instituição científica a funcionar. Matt Taylor geriu a nossa comunidade online e foi um defensor da ciência aberta e da educação científica. Avançou a nossa ciência de formas surpreendentes. Por exemplo, incentivou-nos a transmitir em direto as nossas reuniões internas de investigação, o que, tanto quanto sei, foi uma estreia. O acesso à investigação científica deve ser gratuito. Gostaria de agradecer à SciHub.org, uma organização que

oferece acesso a publicações científicas para quem não tem meios para as pagar.

Donna Dubinsky não é cientista nem engenheira, mas a sua contribuição é insuperável. Trabalhámos juntos há quase trinta anos. A Donna foi CEO da Palm, CEO da Handspring, presidente do Redwood Neuroscience Institute, e é atualmente CEO da Numenta. Quando nos conhecemos, eu tentava convencê-la a aceitar o cargo de CEO da Palm. Antes de ela tomar a decisão, disse-lhe que a minha paixão verdadeira era a teoria do cérebro, e que a Palm era um meio para atingir esse fim. Portanto, dentro de alguns anos, procuraria uma oportunidade para deixar a Palm. Qualquer outra pessoa poderia ter-se afastado nesse momento ou exigido um compromisso indefinido com a empresa. Mas a Donna fez da minha missão a sua missão. Enquanto liderava a Palm, dizia frequentemente aos funcionários que a empresa precisava de ser bem-sucedida para que eu pudesse perseguir a minha paixão pela teoria do cérebro. Não é exagero dizer que nenhum dos sucessos que tivemos na computação móvel, nem qualquer dos avanços científicos na Numenta, teria acontecido se a Donna não tivesse abraçado a minha missão desde o primeiro dia.

O Livro

Demorei dezoito meses a escrever este livro. Todos os dias chegava ao escritório por volta das 7h da manhã e escrevia até às 10h. Embora a escrita seja, por natureza, uma atividade solitária, tive uma companheira e treinadora ao longo de todo o processo: Christy Maver, a nossa vice-presidente de marketing. Apesar de não ter experiência prévia em escrita de livros, aprendeu no terreno e tornou-se indispensável. Desenvolveu um olho clínico para perceber onde eu precisava de dizer menos ou mais. Ajudou-me a organizar o processo de escrita e liderou sessões de revisão do livro com os nossos colaboradores. Embora tenha sido eu a escrever o livro, a presença dela está por todo o lado. Eric Henney, o meu editor na Basic Books, e Elizabeth Dana, a revisora de texto, fizeram inúmeras sugestões que melhoraram a clareza e legibilidade do texto. James Levine é o meu agente literário. Recomendo-o sem reservas.

Quero agradecer ao Dr. Richard Dawkins pelo prefácio encantador e generoso. As suas ideias sobre genes e memes tiveram um profundo impacto na minha visão do mundo, e sou-lhe grato por isso. Se pudesse escolher uma pessoa para escrever o prefácio, seria ele. Sinto-me honrado por o ter feito.

Janet Strauss, a minha companheira de vida, leu os capítulos à medida que os escrevia. Fiz várias alterações estruturais com base nas suas sugestões. Mas mais importante do que isso, ela tem sido a parceira perfeita na minha jornada pela vida. Juntos, decidimos propagar os nossos genes. O resultado, as nossas filhas Kate e Anne, tornaram a nossa breve estadia neste mundo indescritivelmente feliz.

Acerca do Autor

© Fotografia de Tri Nguyen / Tri Nguyen

Jeff Hawkins é cofundador da Numenta, uma empresa de investigação em neurociências, fundador do Instituto de Neurociências Redwood e um dos fundadores da área da computação portátil. É membro da Academia Nacional de Engenharia e autor de "*Sobre Inteligência*".