UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Filosofia e Ciências Humanas

CIRO HIDEKI ARTIGA WATANABE

A TEORIA COMPUTACIONAL DA MENTE E O DILEMA DE SEARLE

CAMPINAS
2021

Ciro Hideki Artiga Watanabe

A TEORIA COMPUTACIONAL DA MENTE E O DILEMA DE SEARLE

Dissertação apresentada ao Instituto de Filosofia e Ciências Humanas da Universidade Estadual de Campinas como parte dos requisitos exigidos para obtenção do título de Mestre em Filosofia.

Orientadora: PROFA. DRA. ITALA MARIA LOFFREDO D'OTTAVIANO

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO Ciro Hideki Artiga Watanabe, ORIENTADO PELA PROFA. DRA. Itala Maria Loffredo D'Ottaviano

CAMPINAS
2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Filosofia e Ciências Humanas
Cecília Maria Jorge Nicolau - CRB 8/3387

W29t

Watanabe, Ciro Hideki Artiga, 1992-
A teoria computacional da mente e o dilema de Searle / Ciro Hideki Artiga Watanabe. – Campinas, SP : [s.n.], 2021.

Orientador: Itala Maria Loffredo D'Ottaviano.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Filosofia e Ciências Humanas.

1. Searle, John R., 1932-. 2. Filosofia da mente. 3. Filosofia e ciência cognitiva. 4. Computação. 5. Metafísica. 6. Cognição. I. D'Ottaviano, Itala Maria Loffredo, 1944-. II. Universidade Estadual de Campinas. Instituto de Filosofia e Ciências Humanas. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: The computational theory of mind and Searle's dilemma
Palavras-chave em inglês:
Philosophy of mind
Philosophy and cognitive science
Computation
Metaphysics
Cognition
Área de concentração: Filosofia
Titulação: Mestre em Filosofia
Banca examinadora:
Itala Maria Loffredo D'Ottaviano [Orientador]
Emiliano Boccardi
Mariana Claudia Broens
Data de defesa: 15-10-2021
Programa de Pós-Graduação: Filosofia

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-9515-7857
- Currículo Lattes do autor: http://lattes.cnpq.br/0240051384521465

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITULO DE FILOSOFIA E CIÊNCIAS HUMANAS

A Comissão Julgadora dos trabalhos de Defesa de Dissertação de Mestrado, composta pelos Professores Doutores a seguir descritos, em sessão pública realizada em 15/10/2021, considerou o candidato Ciro Hideki Artiga Watanabe aprovado.

Profa. Dra. Itala Maria Loffredo D'Ottaviano
Prof. Dr. Emiliano Boccardi
Profa. Dra. Mariana Claudia Broens

A Ata de Defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de fluxo de Dissertações/Teses e na Coordenadoria do Programa de Pós-Graduação em Filosofia do Instituto de Filosofia e Ciências Humanas.

Dedico aos meus pais,
Helena e Roberto.

AGRADECIMENTOS

Este trabalho só foi possível graças ao esforço e boa vontade de muitas pessoas que me ajudaram de maneira direta, lendo e revisando o meu texto, fazendo ilustrações ou discutindo as minhas ideias, ou, de maneira indireta, providenciando as circunstâncias adequadas para o pleno desenvolvimento do projeto.

Expresso meus sinceros agradecimentos à minha orientadora, Profa. Dra. Itala Maria Loffredo D'Ottaviano, amante do conhecimento, que muito me ensinou a respeito do trabalho de pesquisador, e nunca mediu esforços para cumprir sua função de tutora no decorrer dessa custosa, porém, gratificante, empreitada.

Aos professores Emiliano Boccardi e Mariana Broens, que tiveram enorme influência na forma e conteúdo finais deste trabalho. Irei me lembrar com carinho os ensinamentos que me passaram, não apenas durante o processo de mestrado, mas também durante os anos anteriores a ele.

Aos meus colegas do grupo Aleph, que me ajudaram e me incentivaram a iniciar a pós-graduação nesse tema específico e, também, ao professor Marcos Ruffino que, além de me apresentar à profa. Itala, foi presente e solícito em diversas ocasiões.

À Mariana Prete, que não apenas produziu as ilustrações presentes no decorrer do trabalho, mas esteve ao meu lado nos momentos bons e nos momentos difíceis.

Finalmente, aos meus pais, que sempre ofereceram apoio incondicional e me ensinaram mais do que pode ser expresso nessas poucas linhas de agradecimento.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – (CNPq), número de processo 130707/2019-8.

RESUMO

A hipótese de que o cérebro é um sistema processador de informações compõe a base de uma das abordagens teóricas dominantes nas ciências cognitivas hoje. Contudo, essa hipótese sofre ataques de diversas frentes, tanto no campo teórico filosófico como no campo da ciência empírica. Neste trabalho, pretendemos apresentar essa abordagem teórica – conhecida como a Teoria Computacional da Mente (TCM) – e defende-la de um dos ataques de John Searle, filósofo e crítico dessa visão. Além disso, ao final, situamos a TCM com referência às novas teorias que surgiram a partir da década de 1980, viz., os modelos de processamento distribuído paralelo e a cognição incorporada e situada. Em geral, essas alternativas teóricas são vistas como adversárias da TCM. Porém, consideramos a possibilidade de que elas sejam consistentes entre si, de modo que cada teoria possua seu próprio escopo.

Palavras-chave: filosofia da mente; ciência cognitiva; computação; metafísica; cognição.

ABSTRACT

The hypothesis of the brain as an information processing system constitutes the base of one of the dominant theoretical approaches in the cognitive sciences today. However, this hypothesis receives multiple attacks, both from theoretical and philosophical ground and from the empirical sciences. In this work, we intend to present this theoretical approach – known as Computational Theory of Mind (CTM) – and defend it from one of John Searle's attacks, who is a philosopher and critic of the CTM. Also, at the end, we situate the theory by reference to the new approaches that appeared in the decade of 1980, viz., the Parallel Distributed Processing models and the Embodied Cognition. Generally, these alternatives are seen as opponents to the CTM. However, we consider the possibility of treating them as consistent with each other, in the sense that they cover different cognitive domains.

Key-words: philosophy of mind; cognitive science; computation; metaphysics; cognition.

SUMÁRIO

INTRODUÇÃO

Questões acerca da natureza da mente são tão antigas quanto a própria história da filosofia. Elas se dividem em duas categorias, as ontológicas e as epistemológicas. Os problemas ontológicos, ou metafísicos, discorrem sobre o ser da mente ou espírito, e perguntas como "O que é a mente?", "O que é um evento mental?" e "O que é uma propriedade mental"? são levantadas. As tentativas de solucionar os problemas ontológicos, por sua vez, se dividem em duas, as tradições dualista e materialista. O dualismo defende a ideia de que possuímos uma alma distinta do corpo e independente dele, a qual abriga nossos estados conscientes, nossos sentimentos e emoções. Alguns diriam que essa alma é imortal, e que resistiria mesmo após nossos corpos perecerem. A ideia central, porém, é a de que o espírito é de um tipo de substância diferente da substância física, material, que compõe todo o resto do mundo. O mundo, de acordo com o dualista, é dividido em dois reinos, o espiritual e o material, e teria dois tipos de propriedades, as propriedades mentais e as propriedades materiais. Uma pedra, composta de matéria, não sente dor ou prazer, não deseja nada e também não faz inferências lógicas. Do mesmo modo, nossos estados mentais não possuem os predicados que geralmente atribuímos aos objetos materiais. Não dizemos que nossa ideia da Lua tem comprimento x ou que nossa lembrança de um bolo tem peso y. A tradição materialista, por outro lado, nega a existência do reino espiritual, afirmando que tudo o que há pertence ao mundo físico. O que chamamos de eventos mentais são, de acordo com o materialismo, eventos físicos que ocorrem em algum lugar do cérebro[1].

As questões epistemológicas são do tipo "Como representamos ideias e conceitos?", "Como somos capazes de realizar inferências válidas?" e "Quais são as fontes de conhecimento?", dentre outras. Novamente, as respostas para essas questões também se dividem em duas principais tradições filosóficas, o racionalismo e o empirismo. O empirismo é a doutrina segundo a qual todo o nosso conhecimento advém da experiência, ou, de outro modo, que nós nascemos como uma tela em branco, sobre a qual a experiência "pinta" ideias e conceitos. Segundo essa doutrina, nem todas as nossas ideias correspondem a algo que já foi experienciado. Mas, necessariamente, ideias "novas" serão compostas de conceitos mais básicos que, esses sim, foram obtidos através da experiência. A experiência é, portanto, ao mesmo tempo a fonte e, em

[1] Contemporaneamente, os termos *materialismo* e *fisicalismo* são intercambiáveis, embora possuam etimologias distintas (ver Stoljar 2021). Hoje em dia, ambos os termos designam a ideia de que tudo no mundo é físico ou supervém ao físico. Alguns filósofos preferem utilizar o termo *fisicalismo* porque desejam se comprometer com a existência de objetos que não são classicamente tratados como matéria (campos eletromagnéticos, força gravitacional etc.).

certo sentido, a limitadora do conhecimento humano. O racionalismo, por outro lado, argumenta que ao menos algumas ideias são inatas, ou, de outro modo, que nem todo o saber está justificado na experiência. Os racionalistas geralmente tentam mostrar casos em que nosso conhecimento ultrapassa aquele fornecido pelos sentidos, e exemplos de proposições que são conhecidas através de algum tipo de *insight* do intelecto.

Todas essas questões foram amplamente discutidas durante milênios, desde Platão e Aristóteles, passando por Descartes, Locke, Hume, Kant, entre outros, até os dias de hoje (ver Gardner (1985) para uma exposição histórica detalhada). Em nossa história recente (séc. XX em diante), o materialismo tem ganhado cada vez mais adeptos. O dualismo cartesiano, que defende a existência de dois tipos diferentes de substâncias, tem dificuldade em explicar como nosso corpo, uma entidade física, interage causalmente com a nossa mente, uma entidade não-física. Um dos obstáculos frequentemente mencionado nesse contexto é que essa suposta interação causal entre o mundo físico e o mundo não-físico parece violar princípios de conservação de energia e momento (Heil (2013)), e é razoável buscar uma teoria sobre a mente que seja contínua e harmoniosa com nossas ciências mais básicas.

Dado esse cenário cientificamente inspirado, psicólogos buscaram uma teoria cujos princípios fossem compatíveis com os pilares metodológicos da ciência em geral. Assim, era um princípio metodológico convincente, no início do século XX, que o comportamento observável era a única fonte de evidência confiável a respeito de hipóteses sobre eventos e processos mentais de seres humanos e outros animais. Dessa perspectiva, não há diferença alguma entre dois estados mentais que não revelem diferenças comportamentais discrimináveis. Essa abordagem ficou conhecida como *behaviorismo metodológico* (Graham (2019)). Gradativamente, ela se tornou cada vez mais radical, com uma roupagem cada vez mais materialista.

Na década de 1920, John B. Watson [1878 – 1958] influentemente sugere que o modo como nos comportamos é função dos estímulos externos que recebemos do ambiente – sem a intervenção de estados psicológicos internos. A pergunta de como a mente interage causalmente com o corpo é radicalmente cortada na raiz da teoria. Com efeito, qualquer questão a respeito de eventos ou processos mentais é jogada para de baixo do tapete. O domínio da psicologia circunscreve o estudo do comportamento e o histórico causal do organismo com o ambiente. A tarefa dessa ciência consiste em entender como organismos formam associações entre condições externas e respostas comportamentais, como eles são condicionados a responder de uma maneira se o ambiente for de um jeito, e de outra se o ambiente for alterado. Em suma, a psicologia se concentrava em descobrir e descrever as regularidades que expressam o modo como o ambiente controla o comportamento observável. O radicalismo da teoria é expresso na

ideia de que, uma vez que o comportamento é uma função de estímulos externos, qualquer explicação a respeito de como agimos prescinde de termos mentalísticos como *crença*, *desejo*, *expectativa* etc. Essa ideia é extremamente contra intuitiva para o senso comum, uma vez que estamos habituados, há milênios, a explicar as ações dos seres humanos fazendo referência ao que eles acreditam, desejam e esperam dos outros e de si mesmos. Dizemos que João pegou o guarda-chuva porque ele acredita que irá chover, e porque ele não quer se molhar. Ou que Sara bebeu o suco de laranja que estava na geladeira porque estava com sede, e assim por diante. Essa vertente behaviorista ficou conhecida como *behaviorismo radical*. Outro defensor notável dessa doutrina foi B. F. Skinner [1904 – 1990].

Entretanto, as teorias que aplicam os princípios do behaviorismo radical se mostram notavelmente improdutivas, principalmente porque elas rejeitam modelos de explicação em que há causação direta entre estados mentais e comportamento (Fodor 1981). Por isso, surge uma nova forma de behaviorismo, o *behaviorismo lógico*. O behaviorismo lógico é uma teoria semântica a respeito do significado de termos mentalísticos. Dizer que João acredita que irá chover é equivalente semanticamente a dizer que, se João for sair de casa, então ele levará um guarda-chuva. Dizer que Sara está com sede é dizer que, se houver suco de laranja na geladeira, então ela irá beber. As cláusulas se-então são chamadas de *hipotéticos comportamentais*. É possível, diria o behaviorista lógico, e até provável, que o conteúdo semântico de uma proposição que atribui um estado mental a um sujeito seja equivalente a mais de um hipotético comportamental. Assim, *João acredita que irá chover* também significa que se as janelas estiverem abertas, João irá fechá-las, se houver roupas no varal, João irá recolhê-las etc.

A ideia de parafrasear proposições mentalísticas em termos de hipotéticos comportamentais é plausível se pensarmos no modo como aprendemos o significado dessas proposições e suas condições de uso (Graham (2019)). Em condições normais, não diríamos que João acredita que irá chover se ele não estivesse disposto a carregar um guarda-chuva, fechar as janelas e recolher as roupas do varal. Há uma conexão íntima entre a atribuição dessa crença e as disposições comportamentais de João. Além disso, assim como o behaviorismo radical, o behaviorismo lógico oferece uma alternativa teórica que evita algum tipo de dualismo que poderia se seguir caso reconhecêssemos a independência e autossuficiência de termos mentalísticos com relação a estímulos e comportamento aberto, com a vantagem de que o behaviorismo lógico parece acomodar causação entre estados mentais e comportamento. Causação mental nada mais é do que a manifestação de uma disposição comportamental. Ela ocorre quando há a disposição e, além disso, o antecedente do hipotético comportamental é verdadeiro (Fodor (1981)).

Entretanto, o behaviorismo lógico também possui uma fraqueza devastadora. Embora dê conta, a princípio, de acomodar e explicar causação entre estados mentais e comportamento, ele sofre para explicar causação entre estados mentais e outros estados mentais. É extremamente comum que o comportamento seja função da interação entre vários estados mentais. Isto é, João só fechará as janelas se ele acreditar que está chovendo, mas também desejar que o quarto fique seco, tiver a expectativa de que a janela proteja a entrada de água, e assim por diante. Esse tipo de causação o behaviorismo lógico falha em acomodar, porque é difícil explicar como estados mentais, sendo meras disposições, interagem causalmente entre si.

Talvez o principal equívoco do behaviorismo seja o de ignorar as atividades internas do indivíduo, sob a premissa de que o comportamento aberto pode, de fato, ser explicado fundamentalmente com referência ao histórico de estímulos ao qual ele foi exposto. Entretanto, Skinner (1953) acredita que essas atividades internas também sejam 'comportamentos' do organismo – embora comportamentos não visíveis. Dessa perspectiva skinneriana, a tentativa de explicar o comportamento aberto com referência ao comportamento interno do indivíduo incorre em uma espécie de circularidade, pois pressupõe aquilo mesmo que carece de explicação. O argumento é que para explicar o comportamento de maneira não regressiva – ou seja, para não explicar o comportamento com base em outro comportamento – faz-se necessário apelar para elementos que eles mesmos não sejam comportamentais, como estímulos e reforços do ambiente externo. Entretanto, muitas evidências apontavam para a necessidade de levar em conta essas atividades internas. Experimentos revelaram que o apoio exclusivo no histórico de estímulos ambientais era insuficiente para dar conta da versatilidade de nosso comportamento (humano e não humano).

Por exemplo, Edward Tolman (1948) reporta experimentos sobre mapas cognitivos em ratos, os quais buscam caracterizar o que exatamente ratos assimilam quando aprendem a navegar por labirintos. Em um dos experimentos, um grupo de roedores é treinado para percorrer o labirinto da Figura I-1 abaixo. Os ratos partem do ponto A e encontram comida no ponto G. No ponto B eles encontram uma mesa, a qual podem explorar livremente. Eles encontram uma parede que os leva até o ponto D. O ponto H identifica uma luz que ilumina o caminho F-G.

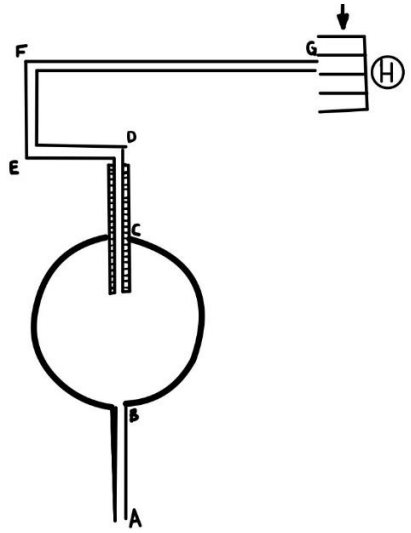

Figura I-1 - Labirinto 1 (adaptado de Tolman (1948)).

Nas primeiras tentativas, os ratos exploram de maneira mais demorada a mesa e o percurso. Tocam e farejam as paredes e, depois de um tempo, se dirigem até a comida. Depois de várias tentativas, os ratos partem do ponto A diretamente para o ponto G. Os behavioristas diriam que os ratos aprenderam a responder da maneira correta aos estímulos que eles recebem. Então, por exemplo, quando eles avistam a mesa, foram condicionados a andar em linha reta e, quando chegam ao ponto D e veem uma curva, foram condicionados a virar à esquerda, e assim por diante, até a comida no ponto G. No entanto, o mesmo grupo de ratos é, posteriormente, testado no labirinto da Figura I-2 abaixo.

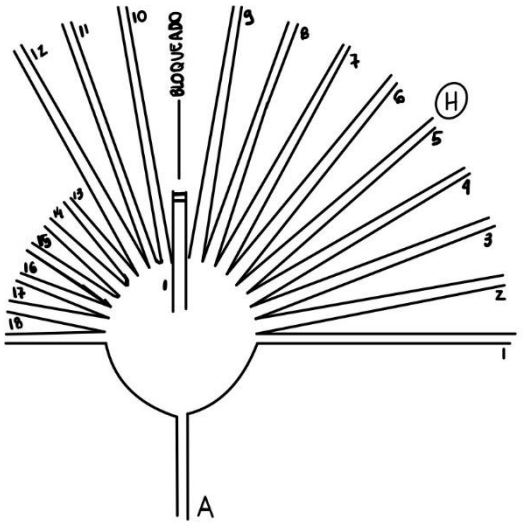

Figura I-2 - Labirinto 2 (adaptado de Tolman (1948)).

Os roedores partem do ponto A e, quando chegam à mesma mesa redonda, encontram o caminho que levaria até o ponto D bloqueado. Eles voltam e exploram cuidadosamente cada um dos novos corredores disponíveis, adentrando apenas alguns centímetros. Após cerca de 7

minutos explorando as possibilidades, a maioria massiva dos ratos escolhe o corredor número 6 – o caminho que leva diretamente à comida –, o que mostra que eles não aprenderam apenas a responder a estímulos específicos, mas adquiriram uma compreensão ampla a respeito do espaço no qual foram sistematicamente treinados. Essa compreensão, evidentemente, consiste em atividades internas do rato, negligenciadas pela corrente behaviorista. Houve, portanto, uma tendência a levar em conta os *processos mentais* que, plausivelmente, figuram como intermediários entre os estímulos e o comportamento observável. Mas o que são esses processos mentais? Teóricos ainda relutavam em abraçar qualquer tipo de dualismo cartesiano que considerasse a mente como algo independente do corpo.

Surge, então, uma alternativa teórica ortogonal em relação às abordagens behavioristas. Na década de 1950, J. J. C. Smart [1920 – 2012], professor na Universidade de Adelaide, argumenta a favor da ideia de que estados mentais devam ser identificados com estados cerebrais. Na época, já havia muitas evidências que correlacionavam eventos neurofisiológicos com eventos psicológicos. Mas a teoria da identidade excede a mera correlação, afirmando que para qualquer evento mental M, haverá um evento neurofisiológico N, tal que M *é* N. Essa hipótese, de acordo com U. T. Place (1956) [1924 – 2000], outro teórico da identidade, deve ser confirmada empiricamente, do mesmo modo que descobrimos de maneira empírica que água é H_2O, ou que nuvens são aglomerados de minúsculas partículas de água, ou que raios não são nada além de descargas elétricas. A ideia é que possuímos conceitos diferentes para o mesmo referente no mundo, mas, com o progresso neurocientífico, seremos capazes de demonstrar que dor, ou prazer, ou a crença de que o Brasil será o campeão em medalhas nas próximas olimpíadas, por exemplo, não são nada além de configurações neurais específicas.

A teoria da identidade supera as dificuldades do behaviorismo apontadas acima. Ela reconhece a existência de estados mentais enquanto entidades reais (e não meras disposições) e pode, portanto, aceitar ao pé da letra as explicações psicológicas que envolvem termos mentalísticos, presentes tanto no senso comum como na psicologia enquanto disciplina acadêmica; ela possui recursos para explicar de que modo estados mentais causam comportamento; e, do mesmo modo, ela possui recursos para acomodar a noção de causação entre estados mentais: uma sequência causal de estados mentais nada mais é do que uma sequência causal de estados neurofisiológicos.

Há dois modos de entender a teoria da identidade. Há a *identidade tipo-tipo* e a *identidade token-token*. A identidade token-token requer apenas que, para um tipo de estado mental M, qualquer instância M_i de M seja idêntica a uma configuração física específica. É

coerente com a identidade token-token que duas instâncias de M, M_1 e M_2, sejam instâncias de configurações físicas de tipos diferentes. A identidade tipo-tipo faz uma asserção mais forte. Ela requer que, para um tipo de estado mental M, qualquer instância M_i de M seja idêntica a um tipo de configuração física específica. Se a identidade tipo-tipo for verdadeira, então a identidade token-token também é, pois se um tipo de estado mental M sempre é instanciado por um tipo neurofisiológico N, então é o caso que todo token M_i é idêntico a uma configuração física específica, a saber, alguma configuração neurofisiológica N_i. Por outro lado, a identidade token-token não implica a identidade tipo-tipo, como foi observado acima.

Essa diferença em força lógica das duas abordagens reflete uma diferença de plausibilidade entre elas. Hilary Putnam [1926 – 2016] oferece um argumento famoso contra a identidade tipo-tipo, favorecendo a ideia de que é empiricamente provável que o mesmo tipo psicológico seja instanciável por mais de um tipo físico (Putnam 1975b). Seu argumento não é conclusivo, embora tenha convencido a maioria dos teóricos da época. Putnam afirma que a tarefa do teórico da identidade tipo-tipo é ambiciosa demais, pois ele precisa mostrar que, para qualquer tipo de estado mental M, há um único tipo de composição físico-química específica FQ, tal que um organismo O instancia M se, e somente se,

a) o cérebro de O possui a estrutura e composição físico-química adequada;
b) O instancia FQ.

Condição Paralela: ao mesmo tempo, FQ não pode ser um estado fisicamente possível de ser instanciado por qualquer organismo que não instancie M.

Note que a, b e Condição Paralela são generalizados para todo estado psicológico. Desse modo, se for possível encontrarmos, para algum organismo não humano, um único predicado psicológico que se aplique a nós e a ele, como 'fome', a identidade tipo-tipo colapsa (assumindo, é claro, que os correlatos físico-químicos da fome nos dois casos sejam diferentes). Ao menos, na época, a maioria dos teóricos acharam a, b e Condição Paralela muitos fortes para serem endossadas, tendo em vista que, no mesmo artigo, Putnam oferece uma opção teórica mais flexível, mas igualmente sólida de um ponto de vista fisicalista.

Por um lado, o teórico da identidade tipo-tipo estava no caminho certo em explicar como estados mentais são causalmente eficientes. Uma instância de estado mental de dor *é* uma instância de estado neurofisiológico no cérebro, e isso explica como o estado mental de dor é causalmente eficiente — não há problema algum (em princípio) em entender como neurônios entram em transações causais com estímulos do ambiente, outros neurônios e com comportamento. Por outro lado, o teórico da identidade tipo-tipo falha em apreciar o caráter relacional de estados mentais (Fodor (1981)). Dor é o estado tipicamente causado por dano

tecidual, que gera o desejo de que ela pare e que causa comportamento aversivo. O que é essencial da propriedade de ser um estado de dor, alguém poderia dizer, é esse conjunto de relações que ela possui com inputs, outputs e outros estados mentais. Esse ponto foi incorporado no behaviorismo lógico com sucesso. O behaviorista lógico estava parcialmente certo em caracterizar a dor como um tipo de disposição, pois, no fundo, a dor é apenas a capacidade de responder a inputs de um certo tipo com outputs específicos (além de se conectar causalmente com outros estados mentais). Da perspectiva behaviorista, não há razões para negar que polvos, por exemplo, sintam dor, pois eles apresentam o mesmo comportamento aversivo que nós frente a danos infligidos em seu corpo. De um ponto de vista psicológico, parece mais importante, para que um estado mental seja atribuído a um organismo, que esse organismo apresente as características relacionais associadas ao estado mental relevante do que esse organismo possuir a constituição físico-química correlacionada aos *nossos* tokens desse estado psicológico. Afirmar o oposto seria defender uma posição "chauvinista" de estados psicológicos (Block (2007)).

Temos, então, que o behaviorista lógico endossa corretamente o caráter relacional de estados mentais, embora falhe em dar conta de seu caráter causal, enquanto que ocorre precisamente o inverso com o teórico da identidade tipo-tipo, o qual satisfaz a dimensão causal dos estados mentais, mas sem levar em conta seu caráter relacional. Putnam avança uma teoria que une os pontos fortes das duas abordagens. De acordo com o filósofo, um estado mental é um estado físico que satisfaz as especificações relacionais (causais) – entre estímulos externos, comportamento e outros estados mentais – características do estado mental relevante. Se dor é o estado mental tipicamente causado por dano tecidual, que causa o desejo de que a dor pare e gera comportamento aversivo, então, o organismo que instanciar um estado físico-químico interno que satisfaça esse papel causal em sua dinâmica psicológica, pode se considerar que ele sinta dor. A capacidade de instanciar uma propriedade mental está associada à capacidade que um organismo possui de instanciar um estado físico interno organizado da maneira apropriada, que cumpra o papel causal que define aquele estado mental. Essa abordagem ficou conhecida como *funcionalismo* sobre a mente, porque ela atribui um papel muito mais importante às relações funcionais de estados psicológicos do que aos substratos físicos que os realizam. O funcionalismo é compatível com a possibilidade de que humanos, polvos e máquinas feitas de silício sejam todos capazes de instanciar dor, basta que instanciem a organização interna apropriada (ver Capítulo 1, §3).

Hoje, o funcionalismo ainda é uma das abordagens dominantes a respeito da ontologia de propriedades mentais. Como era de se esperar, há várias vertentes funcionalistas, as quais variam em força e escopo (ver Levin (2018)). A Teoria Representacional da Mente apresentada

no Capítulo 1 adota o funcionalismo, mas de maneira restritiva. Isto é, o funcionalismo é utilizado para explicar certos aspectos de nossa vida mental, mas alguns limites são reconhecidos. Um importante ingrediente é adicionado a teoria, a noção de símbolo mental. Além de estados mentais como dor e fome, por exemplo, possuímos estados cognitivos de crença, expectativa, dúvida etc. A diferença desses últimos é que eles possuem o que chamamos comumente de conteúdo proposicional. Por exemplo, a sentença "Mariana acredita que a Torre Eiffel fica na França" indica uma relação entre Mariana e a proposição *A Torre Eiffel fica na França*. A Teoria Representacional da Mente se propõe a responder o que significa dizer, *inter alia*, que Mariana possui uma crença, que Mariana possui uma crença com tal e tal conteúdo e como esse estado psicológico se conecta com outros estados psicológicos de Mariana quando ela raciocina. De acordo com a teoria, o funcionalismo é insuficiente para explicar essas noções e, portanto, outros recursos teóricos são recrutados a fim de propor uma possível solução para esses problemas. A noção de uma linguagem do pensamento é crucial nessa empreitada, como veremos brevemente no Capítulo 1. É interessante como questões epistemológicas, a respeito de como o conhecimento é representado e manipulado em nossas mentes, estão entrelaçadas com questões ontológicas da mente.

Contudo, a Hipótese da Linguagem do Pensamento é desafiada por teóricos insatisfeitos com os fundamentos filosóficos da teoria. A principal tarefa desta dissertação é discutir uma das críticas levantadas por John Searle (1992) a esse respeito. Como tarefa secundária, situamos a teoria de Fodor com relação às alternativas contemporâneas que compõem as ciências cognitivas hoje em dia. Para tanto, elaboramos a seguinte estrutura para o texto.

No Capítulo 1, ensaiamos os principais pontos da Teoria Computacional da Mente (TCM) conforme apresentada por Fodor (1975, 1987). A ideia central é a de que a TCM compõe um corpo teórico (mesmo que esquemático) capaz de explicar como o raciocínio ou a cognição ocorre de maneira mecânica. A Hipótese da Linguagem do Pensamento afirma que o cérebro possui um nível de operação em que ele pode ser corretamente caracterizado como processando estruturas simbólicas com propriedades sintáticas. O processamento ocorre inteiramente em virtude das propriedades formais dessas estruturas. Porém, essas propriedades formais 'espelham' as propriedades semânticas dos símbolos, o que garante que as transições entre estados mentais respeitem as relações lógicas entre os seus conteúdos.

No Capítulo 2, expomos detalhadamente a crítica de Searle. Seu principal argumento é o de que a noção de *símbolo* não fixa propriedades intrínsecas de nenhum objeto. Símbolos são convenções humanas e, portanto, não possuem valor explicativo numa teoria da cognição. Uma explicação adequada dos nossos processos mentais, argumenta o autor, deve revelar como o

cérebro *causa* fenômenos mentais, e não é suficiente dizer que é manipulando símbolos, porque símbolos só existem na medida em que são interpretados dessa forma.

No Capítulo 3 discutimos a crítica de Searle, argumentando que o autor interpreta de maneira equivocada a dimensão sintática da teoria. Enquanto que Searle afirma que a TCM pressupõe (ou precisaria pressupor) níveis ontológicos distintos entre as propriedades sintáticas dos símbolos e as propriedades físicas dos veículos que implementam os símbolos, nós argumentamos que isso não é necessário. Consideramos que, de acordo com a TCM, a caracterização formal dos processos mentais constitui um *nível teórico de explicação*, e não implica na existência de um nível ontológico diferente do nível físico. Além disso, oferecemos um contra argumento à tese de Searle de que a realizabilidade múltipla de estados computacionais advém do fato de que esses estados são arbitrariamente atribuídos ao sistema.

O Capítulo 4 situa a TCM com relação às abordagens mais recentes adotadas nas ciências cognitivas, viz., os modelos de processamento distribuído paralelo (PDP) e a cognição incorporada e situada (CIS). O objetivo é apenas expor algumas alternativas ou complementos ao computacionalismo clássico, e não argumentar a favor de uma ou outra tese específica. Os modelos de PDP oferecem uma arquitetura cognitiva alternativa ao computacionalismo clássico, enquanto que a CIS oferece uma visão da cognição como atividade ou processo corporal, e não apenas cerebral.

O Capítulo 5 apresenta as considerações finais deste trabalho. Mantemos a posição de que a TCM, enquanto uma teoria do raciocínio, se destaca entre as demais. Não vemos necessidade, no entanto, de tomar os modelos de PDP e a CIS como teorias completamente inconsistentes com a TCM.

Ao final, o leitor encontra as referências bibliográficas e um apêndice.

1. A MENTE COMPUTACIONAL

O computacionalismo é uma vertente teórica inspirada nas criações de Alan Turing[2] (1936; 1950), e é defendida por uma classe bastante heterogênea de teóricos, como filósofos, linguistas, cientistas e psicólogos cognitivos. Modelos computacionais constituem um importante pilar na compreensão moderna que temos a respeito de diversas áreas da cognição, como aquisição e compreensão da linguagem (Pearl (2010); Kaplan et al (2007); Redington et al (1998)), percepção (Marr (1982); Burge (2010)) e teoria da decisão (Musall et al (2019)). Entretanto, raramente os fundamentos teórico-filosóficos dessa abordagem são explicitados de maneira clara, o que dificulta a própria avaliação dos resultados provenientes de modelagens computacionais nessas áreas. Este capítulo, de caráter expositivo, apresenta os principais pontos do computacionalismo sobre a mente tal como defendido por Jerry Fodor[3] (1975; 1987), em virtude de seu tratamento direto dos fundamentos filosóficos para a teoria e por ser um dos autores mais influentes da Teoria Computacional da Mente. Não pretendemos fazer uma análise crítica dos argumentos a favor da tese, senão apresentá-la de forma sucinta e objetiva.

É importante ressaltar, desde o início, que o computacionalismo de Fodor não é uma teoria do consciente. Isto é, a teoria não afirma que humanos instanciam estados conscientes, com propriedades qualitativas, porque o cérebro implementa determinados processos computacionais. Tampouco a teoria almeja dar conta de habilidades cognitivas como criatividade, flexibilidade e capacidade de adaptação frente a situações adversas. Por último, a teoria não tenta explicar como certos estados mentais possuem conteúdo semântico, mas simplesmente assume isso como um fato. Não obstante, o autor argumenta que o computacionalismo pode explicar alguns processos de alto nível, como processos inferenciais característicos de criaturas inteligentes. O grande mérito da teoria, segundo ele, é explicar como transições entre estados mentais preservam a verdade, ou como passamos de crenças verdadeiras para outras crenças verdadeiras. Em outras palavras, a teoria explica os mecanismos mentais tipicamente empregados no raciocínio.

---

[2] Alan Mathison Turing [1912 – 1954] foi um formidável matemático e um dos pais da ciência da computabilidade. Estudou e formalizou a noção informal de algoritmo (concomitantemente, porém de maneira independente, a outros teóricos, notoriamente Alan Post e Alonzo Church), no que culminou em uma de suas maiores façanhas: a criação da Máquina de Turing, precursora do computador moderno. Sua carreira é também marcada pelos serviços prestados à Inteligência Britânica na Segunda Guerra Mundial, quando chefia uma das seções responsáveis por decodificar mensagens alemãs criptografadas.

[3] Jerry Alan Fodor [1935 – 2017] foi um ilustre filósofo e cientista cognitivo. Lecionou como Professor Emérito na Universidade de Rutgers, Nova Jersey, pela maior parte de sua carreira. Suas contribuições para a filosofia da mente e ciências cognitivas são incontáveis, mas se destacam as teorias da modularidade da mente e da linguagem do pensamento, para citar algumas.

Esses raciocínios são familiares ao leitor. Eles consistem no tipo de racionalidade que expressamos quando inferimos que se está chovendo, João vai tirar as roupas do varal; ou que meu cachorro expressa quando infere que se eu peguei a coleira, ele vai passear. Esse tipo de conhecimento está inserido numa teoria mais abrangente, a psicologia de senso comum. A primeira seção deste capítulo apresenta o que é a psicologia de senso comum. Grosso modo, a psicologia de senso comum (também conhecida como psicologia de crenças e desejos) é uma teoria implícita de conhecimento geral que caracteriza o funcionamento (de parte) da mente humana de acordo com certos princípios e asserções a respeito de estados mentais como crenças, desejos, expectativas etc. (atitudes proposicionais). Fodor defende vigorosamente a psicologia de senso comum.

A Seção 1.2 discorre sobre o ônus explicativo do defensor da psicologia de crenças e desejos. Fodor, enquanto tal defensor, precisa explicar o que são as atitudes proposicionais e revelar os recursos mentais subjacentes às operações que realizamos com elas.

As Seções 1.3 e 1.4 apresentam a perspectiva funcionalista adotada pelo computacionalismo de Fodor. O autor defende apenas uma versão moderada do funcionalismo sobre estados mentais. Ele concorda que crenças e desejos são estados essencialmente funcionais. Ele discorda, porém, da tese de que o conteúdo desses estados possa ser funcionalmente individuado. A Seção 1.4 apresenta também uma ideia central na teoria, que é a noção de diferentes níveis de explicação na compreensão do cérebro.

As seções 1.5 e 1.6 introduzem a hipótese da linguagem do pensamento, pilar fundamental do computacionalismo. Fodor reaviva uma antiga teoria que remete a William de Ockham [4] (1349/1974), e a renova se inspirando em descobertas recentes da lógica contemporânea e da teoria da computabilidade, para explicar como cadeias de pensamento preservam a verdade. A hipótese da linguagem do pensamento, formulada à luz dos insights de Kurt Gödel [1906 – 1978] e Alan Turing, é a chave para explicar mecanicamente como a mente opera quando raciocina.

Por fim, a Seção 1.7 tenta esclarecer a analogia entre mente-cérebro e software-hardware, o que deve ser entendido a partir dela e o que a teoria computacional de fato afirma sobre o cérebro. Veremos que a teoria computacional da mente possui apenas um caráter esquemático, embora estabeleça uma mudança de paradigma fundamental para as ciências cognitivas. Também veremos que ela abre margem para diversas críticas, as quais serão exploradas no próximo capítulo.

[4] 1250 – 1347. Não se sabe exatamente o dia e mês de nascimento e óbito.

1.1 A PSICOLOGIA DE SENSO COMUM

> [...] if commonsense intentional psychology really were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species; if we're that wrong about the mind, then that's the wrongest we've ever been about anything (FODOR, 1988, p. xii).

A principal tarefa de uma disciplina científica é buscar explicações a respeito dos eventos que pertencem ao seu domínio de investigação. Uma vez que diferentes disciplinas se debruçam sobre diferentes domínios do conhecimento, alguém poderia sugerir que diferentes disciplinas estudam diferentes tipos de eventos, como se eventos físicos fossem de tipo diferente de eventos químicos, os quais, por sua vez, seriam diferentes de eventos psicológicos. Porém, se um evento pertence ou não a um determinado domínio de investigação depende em grande medida do modo como ele é caracterizado. Uma sequência de eventos na qual um jogador de basquete arremessa a bola em direção à cesta e erra pode ser analisada de uma perspectiva psicológica, bem como de uma perspectiva física. Um psicólogo e um físico teriam maneiras diferentes de caracterizar *o que* aconteceu precisamente e, como consequência, explicações diferentes de *por que* esse evento ocorreu dessa forma. As ciências empregam estruturas conceituais distintas, o que as levam a descrever os fenômenos à sua própria maneira. Uma análise psicológica de *por que* o jogador de basquete erra quando arremessa a bola não aceita explicações baseadas no ângulo e força de arremesso. Em certo sentido, é óbvio que o ângulo e força de arremesso precisam ser ajustados, mas responder isso ao psicólogo seria não apenas errar a resposta, seria falhar em entender o fenômeno que requer uma explicação. O fato é que tanto o psicólogo quanto o físico podem estudar os mesmos eventos. A diferença é que eles o fazem apoiados em estruturas conceituais diferentes.

Como Pylyshyn (1984, p. 3) argumenta, explicar é uma atividade cujo sucesso depende, em grande parte, de como a explicação se relaciona com o modo como o problema foi concebido. De forma que a estrutura conceitual que é utilizada para construir o problema restringe o tipo de resposta que pode ser oferecida como explicação[5]. Além disso, estruturas conceituais diferentes revelam *generalizações diferentes*, devido ao fato de estruturas conceituais diferentes agruparem eventos de acordo com métricas e parâmetros distintos. Por conseguinte, generalizações diferentes nos levam a *predições distintas*. Isso ocorre porque estruturas

[5] "There are general reasons why one account of a sequence of events might qualify as an explanation while another *true* account of the same sequence does not. These reasons have to do with the fact that such claims as "The occurrence of X (together with...) *explains* the occurrence of Y" are not, in general, equivalent (that is, they need not preserve truth values) when we replace the X or the Y by phrases that refer to exactly the same event or to the same objects". (Pylyshyn 1984, p. 4).

conceituais diferentes capturam padrões diferentes com relação a uma mesma sucessão de eventos.

O objetivo desta seção é mostrar que a psicologia de crenças e desejos, com seus princípios e estrutura conceitual próprias, autoriza-nos fazer generalizações invisíveis a outras abordagens teóricas, e que o comportamento humano só pode ser compreendido através dessas generalizações. Formulação alternativa: o comportamento humano exige explicações pautadas nos termos da psicologia de crenças e desejos, e essas explicações são irredutíveis às explicações encontradas em outras ciências.

Passamos a descrever, brevemente, o que entendemos por psicologia de crenças e desejos (ou psicologia de senso comum).

Temos uma compreensão implícita a respeito dos tipos de estados mentais que humanos instanciam (sentimentos e emoções complexas, estados perceptivos, estados reflexivos) bem como da forma com que eles interagem entre si. Temos uma compreensão implícita a respeito de como esses estados se conectam com estímulos externos e o modo como nos comportamos. Sem perceber, usamos esse conhecimento tácito a respeito da mente humana para navegar pelo mundo social. Usamo-lo tanto porque ele funciona para explicar, em retrospectiva, as atitudes e idiossincrasias do outro, quanto para antecipar suas condutas e comportamentos. Além disso, parece claro que parte da razão pela qual atribuímos toda essa vida mental aos que pertencem à nossa espécie consiste no fato de que projetamos neles, a partir de nossa própria experiência subjetiva, essa variedade de estados internos unidos por uma rede intricada que define suas associações causais. Na literatura, esse conhecimento implícito é chamado de psicologia de senso comum (*folk psychology*).

Apesar de seu caráter implícito, a psicologia de senso comum se assemelha, em muitos aspectos, a uma teoria científica. Como qualquer teoria científica, ela introduz termos teóricos (*crença*, *desejo*, *expectativa*, *medo* etc. (atitudes proposicionais)); ela introduz princípios que articulam esses termos (se S deseja P, e S sabe que a ação A gera o resultado P, então, *ceteris paribus*[6], S age conforme A; ou se S teme que P, então, *ceteris paribus*[7], S deseja que ¬P; etc.); e fornece um corpo explicativo e preditivo próprio, baseado numa taxonomia e princípios próprios. Vale notar que nem todos os autores concordam que a psicologia de senso comum seja um tipo de *teoria* (ver Bermúdez (2005)), mas nós assumiremos que ela é, ou, ao menos, que ela pode ser tratada como tal. Dificilmente ela é o tipo de teoria que pode ser avaliada

[6] Salvas as situações em que S deseje que ¬A, ou que ¬A se sobreponha sobre P nas prioridades de *S*.
[7] Salvas as situações em que S possua razões independentes para tomar ¬P como indesejável também.

empiricamente. O senso comum não diz o que são crenças e desejos, mas simplesmente assume que eles existem. O senso comum é vago demais e implícito demais para que um cientista possa testar seus princípios. Em minha interpretação, o que Jerry Fodor faz pode ser entendido como uma elevação da psicologia de senso comum a um nível de detalhamento que se torna factível, em princípio, avaliar a teoria à luz de evidências das ciências do cérebro e da cognição.

Para voltar ao objetivo desta seção, queremos ilustrar a ideia — defendida por Fodor (1975, 1987), Pylyshyn (1984), dentre outros — de que a psicologia de senso comum constitui um nível de descrição imprescindível na compreensão do comportamento e vida mental humana. Considere a seguinte ficção.

A humanidade construiu a mais poderosa inteligência artificial já vista, o Heuristically Built Algorithm (HAL9000). Na verdade, HAL não foi construído diretamente por humanos, mas é produto do trabalho conjunto dos mais avançados computadores já criados. Seu poder de processamento é digno dos mais fantasiosos contos de Asimov e, devido às já avançadas tecnologias quânticas, seu processador central ocupava o diminuto espaço de uma bola de golfe. A título de comparação com os humanos, todos os grandes mestres enxadristas que jogaram com HAL abandonaram o jogo com menos de 13 lances.

Certa vez, foi dada a HAL a tarefa de prever ações humanas em contextos típicos do cotidiano. A situação apresentada era interrompida pela metade, e HAL precisava responder o que aconteceria na cena seguinte. Por exemplo: Um homem chega em casa e, apertando o interruptor da sala, notava que a lâmpada havia queimado; ele busca na garagem uma escada e uma lâmpada nova, e desrosqueia a lâmpada queimada. "Qual seria sua próxima ação?", era perguntado a HAL. "Não há informações suficientes para responder" – para a surpresa dos cientistas. Outra situação-problema era a seguinte. Uma mulher sai do banho. Ela se veste e pega um par de sapatos em seu armário. Senta-se na beira da cama e calça um dos sapatos. "Qual seria sua próxima ação?", era perguntado a HAL. "Não há informações suficientes para responder". Os cientistas ficaram estupefatos. Nada disso fazia sentido. HAL solicitava informações aparentemente aleatórias, como as condições climáticas da sala e o coeficiente de elasticidade do material do qual eram feitos os sapatos da mulher. Então, os cientistas notaram, após algum tempo, que HAL precisava de uma descrição física detalhada da situação-problema para que ele pudesse simular – a partir das leis físicas que lhe foram implantadas na memória, aplicadas ao estado de coisas em que a cena fora cortada – qual seria o estado da situação seguinte.

Essa ficção tem o objetivo de ilustrar o seguinte ponto: HAL era incapaz de prever (ou entender) o comportamento dos agentes da história porque ele via o mundo a partir de uma

perspectiva física (empregando a estrutura conceitual e os princípios da física). Nós, humanos, conseguimos prever (e entender) o comportamento dos agentes da história porque vemos o mundo a partir de uma perspectiva psicológica (empregando a estrutura conceitual e os princípios da psicologia de senso comum). É só entendendo o modo como os agentes das situações-problema percebem e representam o mundo, quais são suas intenções, crenças e desejos, que podemos fazer previsões e fornecer explicações confiáveis de suas ações. HAL, no entanto, não estava equipado com esse repertório. O único tipo de descrição que ele era capaz de atribuir às situações era físico, não psicológico, e por isso ele falhava em completar a tarefa. Somos capazes de antecipar as ações das personagens da história porque reconhecemos padrões cognitivos[8]. Não é preciso conhecer a constituição física da lâmpada (se é fluorescente ou incandescente), da escada (se é de alumínio ou de ferro), ou da altura do teto. O padrão cognitivo é evidenciado independentemente desses detalhes, pois existem potencialmente infinitas situações onde uma pessoa usa uma escada para trocar uma lâmpada, e cada uma delas possuirá uma descrição física diferente. Essas descrições físicas, no entanto, não são condições necessárias e nem suficientes para capturar o que existe de comum entre todas elas: o fato de que são situações onde alguém usa uma escada para trocar uma lâmpada. Nós conseguimos classificar a ação do sujeito como uma instância de "trocar uma lâmpada" porque reconhecemos, dado o modo como a situação foi descrita, que o sujeito *percebe* que a lâmpada da sala estava queimada, que ele *deseja* que a sala fique iluminada, que ele *sabe* que ele precisa da escada e de uma lâmpada nova e que ele *se lembra* que esses objetos ficam guardados na garagem.

HAL é capaz de descrever a situação apenas de uma perspectiva física e, portanto, ele utiliza noções como massa, temperatura, força, pressão etc., para caracterizar a sequência de eventos que ele observa. A situação aqui é similar àquela em que um jogador de basquete arremessa a bola em direção à cesta: existe, de fato, uma descrição física correta acerca da sequência de eventos que constitui a cena; no entanto, adivinhar a próxima ação do indivíduo requer um entendimento a respeito do que ele está fazendo, e o que ele está fazendo não tem relação alguma com a massa da lâmpada, temperatura da sala ou coeficiente de atrito do solo. Isso se torna evidente quando consideramos situações contrafactuais nas quais as condições físicas da cena mudam, mas a explicação e descrição do comportamento do sujeito permanecem as mesmas. Em vez da escada, ele poderia usar uma cadeira, por exemplo. Também poderiam

[8] E também porque fazemos parte do mesmo nicho cultural a que pertencem as personagens da história. Uma pessoa que nasceu no século XV, época em que não havia lâmpadas e rede elétrica, presumivelmente não compreenderia a situação descrita. Mas isso não indica uma incapacidade de capturar os padrões cognitivos relevantes, apenas uma falta de conhecimento do que são lâmpadas e interruptores.

haver situações cujas condições físicas são idênticas, mas a descrição do comportamento do sujeito muda. Ele poderia estar encenando numa peça de teatro, por exemplo.

Qualquer explicação aceitável a respeito do comportamento do indivíduo invocaria razões, motivos e condições nas quais ele se encontra (especificadas em termos psicológicos, e não físicos), de modo que as especificações materiais da lâmpada ou da escada ficam em segundo plano. Qualquer descrição psicológica de uma ação ou comportamento corresponde a indefinidamente muitas descrições físicas, nenhuma das quais é necessária (e, portanto, relevante) para entende-la[9]. É claro que, em certo sentido, a próxima ação do sujeito é um resultado das leis físicas aplicadas ao estado de coisas anteriores. Porém, quando estamos falando de comportamento racional, qualquer explicação em termos físicos soa inadequada, pelo fato de que essa estrutura teórica deixa escapar as generalizações que são importantes. Tais generalizações só são exprimíveis com um vocabulário intencional, pautado nas atitudes proposicionais e nos princípios da psicologia de senso comum. O indivíduo atua conforme as leis da física, mas também segue as leis (princípios da psicologia) que caracterizam o comportamento racional (Pylyshyn, 1984)[10].

Assim, a psicologia de senso comum constitui um nível teórico a partir do qual se descreve e se expressa generalizações legiformes de nível pessoal[11], articulando e empregando um repertório conceitual finamente ajustado para essa tarefa. Nenhum outro nível teórico (físico, químico ou biológico) captura esses padrões e generalizações. Nenhum outro nível teórico parece adequado, portanto, para compreender o comportamento racional de humanos (e alguns animais)[12]. A noção de que uma mesma sequência de eventos admite diversos níveis de caracterização, e que algumas caracterizações são mais adequadas do que outras em função do modo como o explanandum é definido, terá grande importância no decorrer dos capítulos (ver também Apêndice ao final).

---

[9] "The point is, there are innumerably many physically distinct ways in which the same generalization can be realized; yet they remain cases of the same generalization. If that generalization were not recognized, *each instance would count as a different sequence*, and we would miss an important regularity" (Pylyshyn 1984, p. 8).

[10] Pylyshyn (1984) oferece um exemplo instrutivo a esse respeito. Num jogo de beisebol, os jogadores seguem as leis da física, mas também seguem as leis (regras) do jogo. Não há conflito algum aqui. Mas para explicar por que um jogador correu da segunda para a terceira base, é inútil apelar para as leis da física.

[11] O nível pessoal é o nível de organização do indivíduo como um todo, tipicamente aquele em que as generalizações da psicologia de senso comum são expressas. Também temos as generalizações da química e da macro economia, cada uma situada em níveis de organização diferentes.

[12] Esse ponto é amplamente disputado na filosofia da mente. Ver, por exemplo, Churchland (1981) e Dennett (1989). Paul Churchland argumenta que a psicologia de senso comum é uma teoria não apenas estagnada, mas em constante regressão, com grandes chances de ser provada falsa por uma neurociência completa. Entretanto, não caberia aqui uma discussão a respeito desse embate teórico. Nosso objetivo é discutir as críticas de John Searle (ver capítulos 2 e 3), o qual não levanta nenhuma objeção específica à psicologia de senso comum.

1.2 A TEORIA REPRESENTACIONAL DA MENTE

Se aceitarmos a psicologia de crenças e desejos como uma teoria adequada a respeito da mente humana, o próximo passo é saber com o que estamos nos comprometendo, com quais tipos de entidades e com princípios de que natureza. Se este fosse um trabalho sobre filosofia da física, estaríamos nos perguntando "O que são elétrons e neutrinos, e o que são leis físicas?". Ao invés disso, queremos saber que tipo de coisas são crenças, desejos e outras atitudes proposicionais. Também queremos um fundamento teórico a respeito dos mecanismos que garantem as relações sistemáticas expressas pelas generalizações da psicologia de senso comum.

É importante entender, em primeiro lugar, que o tipo de explicação oferecida pela psicologia de crenças e desejos é *causal*. (Ao menos, as explicações de senso comum respeitam relações contrafactuais características de relações causais: e.g. se S não desejasse que D e não acreditasse que C, S não teria agido conforme A). A psicologia de senso comum está comprometida com pelo menos três tipos de causação envolvendo estados mentais: a causação de comportamento por estados mentais; a causação de estados mentais por eventos do ambiente externo que interagem com nossos órgãos sensoriais; e a causação de estados mentais por outros estados mentais (Fodor 1987, p. 12).

Em segundo lugar, é importante notar a dimensão *semântica* das atitudes proposicionais. Quando dizemos que Tom acredita que Bia é mais alta que Matheus, significa que Tom representa o mundo (ou melhor, parte dele) como sendo de uma certa forma, a saber, Tom representa o mundo como sendo tal que Bia é mais alta que Matheus. Quais características a representação de Tom possui? Podemos dizer que a representação mental de Tom é uma espécie de imagem mental na qual Bia aparece mais alta que Matheus? Talvez. O problema com essa ideia é o de que uma imagem mental desse tipo pode expressar mais de uma coisa. Essa imagem pode representar o estado de coisas em que Bia é mais alta que Matheus, mas também o estado de coisas em que Bia e Matheus possuem 2 pernas cada um, ou o estado de coisas em que os dois possuem mãos. Mas, observe que quando atribuímos a Tom a crença de que Bia é mais alta que Matheus, não estamos atribuindo a ele nenhuma crença a respeito da quantidade de pernas ou mãos que Bia e Matheus possuem. Além disso, parece possível que a única crença que Tom possua a respeito de Bia e Matheus é a de que Bia é mais alta que Matheus, e nenhuma imagem parece corresponder a essa ideia simples.

De qualquer forma, Fodor nega que seja esse tipo de representação (imagética) que está envolvido com as atitudes proposicionais. Na próxima seção será discutido que tipo de

representação Fodor acredita estar vinculado com as atitudes. No momento, queremos apenas destacar que há uma dimensão semântica — representacional — envolvida, a qual requer alguma explicação adicional se desejamos tomar a psicologia de senso comum como uma teoria adequada da mente humana.

Assim, aceitar a psicologia de crenças e desejos como uma descrição adequada a respeito da mente humana é aceitar um catálogo ontológico de estados mentais, reconhecendo (a) o papel causal que cumprem em nossa dinâmica psicológica e (b) a dimensão semântica que eles possuem.

De acordo com Fodor (1985), qualquer teoria que incorpore a psicologia de senso comum de maneira sensível às condições (a) e (b) acima recebe o nome de Teoria Representacional da Mente (TRM). Ocorre que temos diferentes Teorias Representacionais da Mente em função de existirem diferentes modos de entender a dimensão causal de nossos estados psicológicos e de existirem diferentes maneiras de caracterizar as propriedades representacionais das atitudes proposicionais. Fodor afirma que o empirismo clássico da Modernidade, por exemplo, é um tipo de Teoria Representacional da Mente (Fodor 1985, p. 286), de acordo com a qual nossas representações (ideias) possuem uma natureza sensível (i.e., são cópias de impressões da percepção), e o papel causal que define as transições entre estados mentais é baseado em mecanismos de associação[13]. De acordo com esse tipo de TRM, portanto, meu pensamento de que está chovendo lá fora envolve uma espécie de simulação mental de experiências sensoriais que tive em momentos anteriores quando percebi, através dos sentidos, que chovia. Sucessões causais de estados mentais, por sua vez, são explicados por associações sedimentadas na experiência. O pensamento de que está chovendo causa o pensamento de que o chão deve estar molhado porque essas duas ideias correspondem a impressões sensoriais que vieram acompanhadas inúmeras vezes na experiência. Essas ideias estão fortemente conectadas por um mecanismo mental de associação.

De acordo com Fodor, porém, o empirismo clássico falha em tentar reduzir os mecanismos de transição entre estados mentais à mecanismos de associação. A característica mais intrigante das transições entre estados mentais é a de que elas possuem, frequentemente, uma *forma argumentativa*. A relação que existe entre um estado mental A que causa um mental B é, muitas vezes, o mesmo tipo de relação que existe entre uma premissa e uma conclusão: uma relação de implicação lógica. Assim, por exemplo, alguém que acredita que (a Torre Eiffel fica

[13] Hume (2007).

em Paris & Paris fica na França), provavelmente também acredita que a Torre Eiffel fica na França, porque essa é uma conclusão lógica da sentença anterior. O associacionismo empirista explica transições entre estados mentais fazendo referência unicamente às sucessões entre itens mentais que correspondem a sucessões que ocorreram na experiência de maneira sistemática[14]. Entretanto, isso não parece suficiente para explicar a relação lógica que as transições entre estados psicológicos muitas vezes respeitam.

Portanto, precisamos de uma teoria que dê conta das interações causais entre nossos estados mentais — levando em consideração as relações lógicas muitas vezes mapeadas por elas — e que discorra sobre o aspecto representacional de nossos estados psicológicos.

1.3 FUNCIONALISMO SOBRE ESTADOS MENTAIS

Nesta seção, veremos como o funcionalismo é incorporado na Teoria Representacional da Mente defendida por Jerry Fodor. Em linhas gerais, a doutrina funcionalista cumpre uma função na individuação das atitudes. Assim, um estado mental é identificado com uma crença apenas se ele cumprir a função que crenças tipicamente cumprem na dinâmica psicológica de um organismo.

Em termos simplistas, o funcionalismo é a ideia de que a classe (tipo) de um objeto funcional deve ser individuada, definida ou reduzida ao conjunto de relações causais exercidas por esse objeto dentro do sistema no qual ele opera. Essa é apenas uma maneira elegante de dizer que objetos funcionais são individuados de acordo com o que eles fazem dentro de um sistema, em oposição a sua composição material. O modo mais direto de entender o funcionalismo é através de exemplos. Artefatos humanos são os casos paradigmáticos de objetos funcionais. Relógios, ampulhetas e termostatos são objetos funcionais porque suas identidades estão amarradas às funções que eles executam. Um relógio pode ter engrenagens internas que coordenam os movimentos de ponteiros, mas também pode ser construído a partir de microprocessadores eletrônicos e um display. O que o torna um relógio é sua capacidade de informar as horas à pessoa que utiliza esse aparelho. Na literatura, dizemos que propriedades funcionais são *multiplamente realizáveis*. Informar as horas é uma propriedade multiplamente realizável porque existe uma gama irrestrita de sistemas com constituição interna inteiramente

[14] Isso está um pouco impreciso. Hume aceita outros tipos de associação entre ideias, como associação por semelhança (Hume, 2007). Porém, isso ainda parece insuficiente para explicar transições lógicas entre estados mentais.

singular e, no entanto, todos executam a mesma função e, portanto, todos pertencem à mesma classe.

A doutrina funcionalista na filosofia da mente defende que tipos mentais são tipos funcionais. Crenças, desejos etc. devem ser individuados de acordo com o papel que cumprem dentro do sistema do qual fazem parte. Em outras palavras, segundo o funcionalismo, um token de estado mental é definido pelas relações causais que mantém com estímulos externos, outros estados mentais e comportamento manifesto (Levin 2018, Lowe 2000). Um exemplo clássico, a título de ilustração, é o estado de dor. Esse estado psicológico é tipicamente (1) causado por dano tecidual, (2) gera o desejo de que a dor pare e (3) causa comportamento aversivo. De acordo com o funcionalismo, um sistema que exiba esse tipo de organização funcional é capaz de experienciar a dor, independentemente de sua constituição biológica. A tese funcionalista pode parecer um pouco duvidosa fora de contexto, mas ela constitui uma resposta muito mais flexível ao problema da individuação de estados mentais do que as teses adversárias. Uma tese alternativa ao funcionalismo é a de identidade tipo-tipo, de acordo com a qual tipos de estados mentais são idênticos a tipos físicos (neurais, no nosso caso). Um tipo de estado mental, como dor, é idêntico a um tipo de configuração neural e, portanto, toda instância de dor é uma instância desse tipo de configuração neural. O funcionalismo é mais flexível do que a identidade tipo-tipo porque, ao mesmo tempo em que o funcionalismo é consistente com o materialismo aspirado pela identidade tipo-tipo, ele acomoda a intuição de que seres com configurações neurais diferentes da nossa podem experienciar o mesmo tipo de dor. Isso porque propriedades mentais são propriedades organizacionais do sistema.

Assim, se estados psicológicos forem (puramente) estados funcionais, seria possível, a princípio, haver os mesmos tipos de processos mentais que os nossos ocorrendo em um sistema de natureza física completamente diferente da nossa, como em um robô feito de silício. Basta que ele replique a mesma organização funcional realizada pelo nosso cérebro, uma vez que o funcionalismo é neutro com relação ao tipo de material que realiza as funções. O funcionalismo se importa apenas com o nível abstrato de organização funcional, o qual é multiplamente realizável.

O funcionalismo, no entanto, também apresenta problemas. Talvez o maior deles seja sua notória incapacidade de lidar com o aspecto qualitativo de nossos estados mentais (Levine 1983). Parece razoável que a individuação de um estado mental deva levar em conta também suas propriedades qualitativas. Por exemplo, se ao olhar para objetos azuis Joana de fato enxerga a cor azul, mas Maria enxerga a cor vermelha, elas obviamente instanciam estados mentais de tipos diferentes, muito embora esses estados mentais sejam indistinguíveis do ponto de vista

funcional (elas apresentarão o mesmo comportamento sempre que virem objetos de cor azul, sempre que perguntadas se um objeto é azul ou não etc.). Assim, precisamos ser cuidadosos ao aceitar o funcionalismo na individuação de estados psicológicos.

A Teoria Representacional da Mente defendida por Fodor incorpora o funcionalismo apenas na individuação de tipos de atitudes proposicionais: crenças, desejos, expectativas etc[15]. Instanciar o estado mental de crença, portanto, não é nada mais do que instanciar um estado funcional tal que esse estado satisfaça o papel teórico definido pelas ocorrências do termo *crença* na rede de generalizações legiformes que caracterizam a psicologia de senso comum em termos das atitudes proposicionais (e.g. uma das generalizações a respeito do estado de crença é que se S sabe que P e que (P → Q), então S sabe que Q. Instanciar o estado de crença envolve instanciar um estado que satisfaça essa e todas as outras generalizações legiformes em que há a ocorrência do termo *crença*). O mesmo vale, *mutatis mutandis*, para as outras atitudes proposicionais.

Assim, a estrutura funcional que caracteriza a mente humana pode ser entendida como uma constelação complexa que interliga estados mentais com outros estados mentais, com estímulos externos e com comportamento. Essa rede, evidentemente, comporta muitos outros tipos de estados psicológicos e processos cognitivos além daqueles especificados e expressáveis através do vocabulário das atitudes. A psicologia de senso comum destaca apenas alguns nós específicos (viz., crença, desejo, expectativa etc.). Cada nó codifica o papel teórico que ocupa no nexo de princípios e generalizações que constitui a psicologia de senso comum. Ocorre que o cérebro, de acordo com o funcionalismo, implementa uma estrutura organizacional isomórfica a essa rede[16].

Entretanto, é importante notar que, para Fodor, apenas os tipos de atitudes são funcionalmente individuados, e não os subtipos. Isto é, propriedades funcionais de estados mentais diferenciam, por exemplo, crenças de desejos e de expectativas (tipos de estados), mas não a crença de que A da crença de que B (subtipos do estado de crença). Ou seja, de acordo com Fodor, uma atitude proposicional não possui o conteúdo que ela possui em razão do papel causal que ela cumpre no sistema ao qual ela pertence. Algumas razões que o autor oferece para

[15] Dor não é uma atitude proposicional. Atitudes proposicionais sempre possuem um conteúdo proposicional. Conteúdo proposicional é aquele que pode, em princípio, ser expresso por uma sentença declarativa e, além disso, pode ser avaliado como verdadeiro ou falso (embora possa haver proposições cujo conteúdo não seja exprimível por uma sentença declarativa do português, mas isso indicaria apenas uma limitação semântica do próprio sistema linguístico). Toda atribuição de atitude proposicional vem acompanhada da atitude (crença, desejo etc.) e do conteúdo proposicional ao qual a atitude se direciona. "Tom acredita que Bia é mais alta que Matheus" é a atribuição da atitude de crença à Tom, cujo conteúdo é *Bia é mais alta que Matheus*.

[16] Ver D'Ottaviano e Filho (2018).

rejeitar essa possibilidade são expostas na Seção 1.5, quando falarmos de como o conteúdo representacional das atitudes proposicionais é estruturado (de acordo com a TRM).

1.4 A PSICOLOGIA E OS NÍVEIS DE EXPLICAÇÃO

Os termos teóricos da psicologia são predicados de indivíduos inteiros. É o sujeito inteiro, enquanto um sistema único que se destaca do meio, que conhece, que age, que teme, que percebe ou possui expectativa; e não, por exemplo, uma parte de seu cérebro. Concomitante a isso, temos uma vasta coleção de especialidades científicas que se situam em níveis diferentes de organização, tanto abaixo quanto acima da psicologia de senso comum. Sociólogas atentam para sistemas compostos de indivíduos, para a dinâmica que emerge da interação entre eles. O vocabulário da socióloga contem predicados que se aplicam a grupos de indivíduos enquanto um sistema único, o que lhe permite evidenciar padrões invisíveis à níveis teóricos hierarquicamente mais abaixo, mesmo sendo o caso que, em alguma medida, a dinâmica social seja uma função do modo como os indivíduos pensam e raciocinam. A mesma ideia se aplica à psicologia de senso comum. Esse nível de abstração descreve o modo como indivíduos pensam e raciocinam e, em alguma medida, o modo como indivíduos pensam e raciocinam é uma função da maneira como cada parte do cérebro opera. Dito isso, não estaríamos sendo precipitados em defender a psicologia de senso comum, se não sabemos detalhadamente como cada parte do cérebro funciona? Não deveríamos esperar a neurociência se desenvolver para que ela nos informe sobre como nós pensamos e raciocinamos de fato? Não seria o caso que as explicações da neurociência tivessem algum tipo de prioridade epistêmica sobre as explicações da psicologia? O que faríamos se encontrássemos incompatibilidades entre os dois níveis de explicação?

Acreditamos não haver essa prioridade epistêmica da neurociência sobre a psicologia, do mesmo modo que, plausivelmente, sociólogos são plenamente capazes de analisar a dinâmica de grupos sociais sem se aprofundar muito em assuntos psicológicos. No caso da relação entre as explicações da psicologia e as explicações da neurociência, é razoável esperar que uma complemente a outra, do seguinte modo. Para entender como um sistema complexo funciona, nem sempre o melhor ponto de partida é analisar o comportamento de suas partes menores para depois tentar entender como elas se unem para compor o todo. Às vezes, a melhor estratégia toma o sentido inverso: deve-se conhecer o comportamento geral do sistema para, depois, investigar como as partes internas se articulam para gera-lo. Nesse sentido, a tarefa de investigar a psique humana se assemelha a um trabalho de engenharia reversa. Suponha que uma espaçonave alienígena caia na Terra, e um grupo de cientistas e engenheiros se encarregue de

entender como a nave funciona. Não parece razoável que a primeira coisa a se fazer seja abrir a nave, separar os seus componentes e estudá-los separadamente. Ao contrário, a primeira coisa a se fazer é, sem desmontá-la, apertar seus botões, puxar suas alavancas e ver o que acontece. O primeiro passo é *conhecer suas funcionalidades*. Apenas depois disso é que se deve seguir os fios e ver quais mecanismos realizam quais funções, e a partir de quais meios. Note que, de maneira análoga ao estudo do cérebro, compreender o funcionamento da nave exigirá uma série de níveis teóricos interdisciplinares: o engenheiro de software se encarregará de entender os controles operacionais; o engenheiro elétrico analisará os circuitos; o engenheiro mecânico se encarregará de analisar como as operações são levadas a cabo mecanicamente; o químico investigará quais recursos materiais a nave consome para manter suas funções; e assim por diante. Nesse sentido, a atividade do psicólogo se assemelha à atividade do engenheiro de software da espaçonave: consiste em desvendar quais funções cognitivas o cérebro é capaz de implementar. Paralelamente, as diversas áreas da neurociência estudam quais mecanismos neurais são recrutados para que essas funções sejam executadas com sucesso[17]. (Ver Apêndice).

Que o sistema nervoso admite uma série de níveis de compreensão, situados em níveis diferentes de organização, é um fato. O que é objeto de disputa (ver Capítulo 2) é se o cérebro admite um nível funcional/computacional de caracterização. A Teoria Representacional da Mente representada por Fodor defende que sim. Essa caracterização se situa entre o nível da psicologia de senso comum e o nível neurofisiológico. Vimos na Seção 1.3 que as atitudes são individuadas de acordo com o papel funcional que cumprem no sistema. Além disso, como veremos nas próximas páginas, algumas operações do cérebro são descritas pela TRM como computacionais, i.e., como operações sobre estruturas simbólicas. A caracterização computacional das operações do cérebro também está em um nível hierarquicamente acima do nível de caracterização neurofisiológico.

1.4.1 MARR E OS TRÊS NÍVEIS DE EXPLICAÇÃO

Um dos primeiros autores a aplicar sistematicamente essa abordagem ao cérebro foi David Marr [1945 – 1980]. Marr (1982) aceita a ideia de que os processos cerebrais possuem diferentes níveis de explicação, e que uma compreensão integral desses processos exige um esforço essencialmente multidisciplinar. Marr se baseia na premissa de que nossas funções cognitivas consistem em densos processamentos de informação. Dessa perspectiva, um sistema

[17] Esse tipo de análise é chamado de *top-down*, notoriamente empregada por Marr (1982) no estudo da visão humana.

(e.g. auditivo, visual etc.) recebe informação de certo tipo, trata esses dados e devolve, como resposta, informação de certo outro tipo.

A primeira coisa que devemos fazer quando estudamos um sistema específico, sugere o autor, é caracterizar suas operações nesses termos, viz., quais informações esse sistema possui como input e quais ele devolve como output. Ou seja, qual função – em termos de tratamento informacional – esse sistema implementa. Marr chama esse primeiro nível de *nível computacional*.

O próximo estágio da investigação consiste em tentar descobrir quais passos computacionais são executados pelo sistema no decorrer do processo. A função cognitiva pode ser dividida em etapas nas quais a informação vai sendo gradativamente refinada e transformada. O *nível algorítmico* especifica as instruções que o sistema é programado para executar tal que os inputs informacionais indicados no nível acima sejam transformados nos outputs com sucesso.

Por último, os investigadores se debruçam sobre o cérebro para entender como as operações especificadas no nível algorítmico são implementadas fisicamente (neurologicamente). O *nível implementacional* consiste em investigar os recursos neurofisiológicos empregados pelo cérebro para que o algoritmo seja executado de maneira concreta.

Nível computacional	Nível algorítmico	Nível implementacional
Qual é a função que define a operação desse sistema cognitivo?	Como essa tarefa computacional pode ser implementada? Quais passos algorítmicos estão envolvidos?	Como o cérebro implementa esse algoritmo fisicamente?

Tabela 1.1 - Os três níveis de investigação propostos por Marr (1982).

É importante observar que, além da proposta dos três níveis distintos de descrição dos processos cerebrais, Marr ao mesmo tempo os integra, formando uma explicação unificada, embora, de algum modo, hierárquica, de nossas operações cognitivas.

O autor emprega essa divisão em sua obra *Vision*, na qual ele expõe os princípios computacionais através dos quais nosso sistema visual opera. Como vimos, o primeiro passo da investigação consiste em descrever qual a função executada por esse sistema. Nessa etapa, Marr formula sua hipótese a partir da observação de pacientes que sofreram lesões no lobo parietal. Dependendo do local em que a lesão ocorreu, diferentes déficits cognitivos irão se manifestar. Pacientes que sofreram lesão no lobo parietal *direito* são capazes de identificar e reconhecer objetos se esses objetos forem apresentados de uma perspectiva típica. Caso o objeto seja

apresentado de uma perspectiva atípica, pacientes não apenas não irão reconhecê-lo, mas negarão que as formas observadas correspondem ao objeto apresentado.

Figura 1.1 - Ao lado esquerdo, uma imagem de um balde capturada de um ponto de vista convencional. À direita, a imagem de um balde capturada de um ponto de vista não convencional (retirada de Bermúdez (2014)).

Por outro lado, pacientes com lesão no lobo parietal *esquerdo* apresentam performance regular no reconhecimento de objetos, tanto de perspectivas típicas como de perspectivas atípicas. O problema com esses pacientes é o de articular verbalmente o nome e função do objeto identificado. A conclusão de Marr é a de que informações a respeito do formato dos objetos são processadas independentemente de informações a respeito de para quê eles servem e como os nomeamos. A função do sistema visual, portanto, é formar uma representação da forma e arranjo espacial dos objetos externos, de modo que essa representação seja centrada nos objetos, em oposição a serem centradas no observador. A ideia é que nosso sistema visual, quando opera normalmente, é capaz de extrapolar as propriedades espaciais dos objetos mesmo se elas estiverem ocultas de nosso ponto de vista. Por isso, a descrição final do objeto que resulta da operação do sistema visual primário foge à perspectiva parcial do observador, centrando essa descrição no próprio item observado.

Quando partimos para o nível algorítmico, essa função computacional precisa ser descrita de maneira muito mais detalhada e, portanto, perguntas como "Como o cérebro representa a informação?", "Quais são os primitivos representacionais sobre os quais o sistema opera?" e "Quais tipos de operações são realizadas?" acabam surgindo. Os detalhes da teoria de Marr são muito complexos, mas a ideia geral é que a representação do objeto externo é processada de maneira gradativa, de modo que cada etapa do processamento fornece um incremento ou melhoria da etapa anterior. Marr chama essas etapas de primeiro esboço (*primal sketch*), esboço 2.5D (*2.5D sketch*) e esboço 3D (*3D sketch*).

Na primeira etapa, nosso sistema visual detecta borrões, linhas virtuais, limites ou bordas do objeto, sobretudo através de variações bruscas na intensidade de luz (chamadas de *zero-*

crossings). Em geral, nessa etapa, é possível se ter uma ideia da geometria básica do objeto observado. A Figura 1.2 abaixo seria um exemplo de primeiro esboço, onde é possível notar a existência de um triângulo que se destaca no centro.

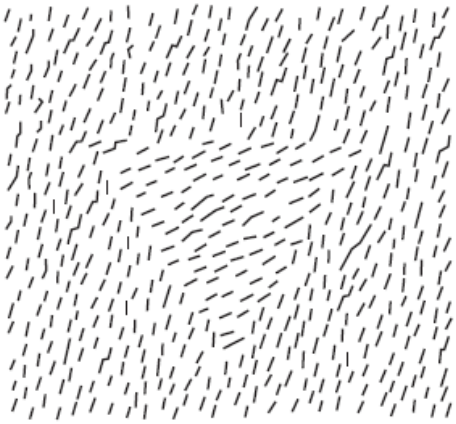

Figura 1.2 - Exemplo de primeiro esboço de nosso sistema visual primário (adaptado de Marr (1982)).

No estágio seguinte de processamento, nosso sistema visual captura informações sobre a orientação e profundidade das superfícies de objetos.

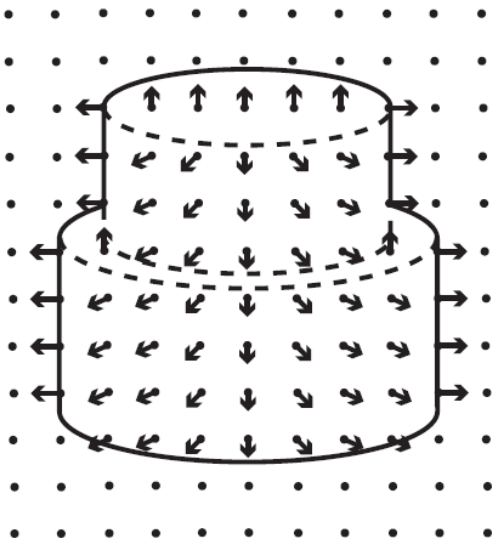

Figura 1.3 - Exemplo de esboço 2.5D de nosso sistema visual (adaptado de Marr (1982)).

Esses dois primeiros esboços, porém, ainda são centrados na perspectiva do observador. O esboço 3D, que constitui o produto final de nosso sistema visual primário, forma uma descrição do ambiente externo centrada nos objetos. Isto é, a representação atinge um estágio de constância perceptual que supera o nosso ponto de vista e as mudanças que ocorrem de nossa

perspectiva caso o objeto esteja em movimento em relação a nós. Essa representação centrada no objeto nos permite reconhecer um mesmo item de várias perspectivas diferentes.

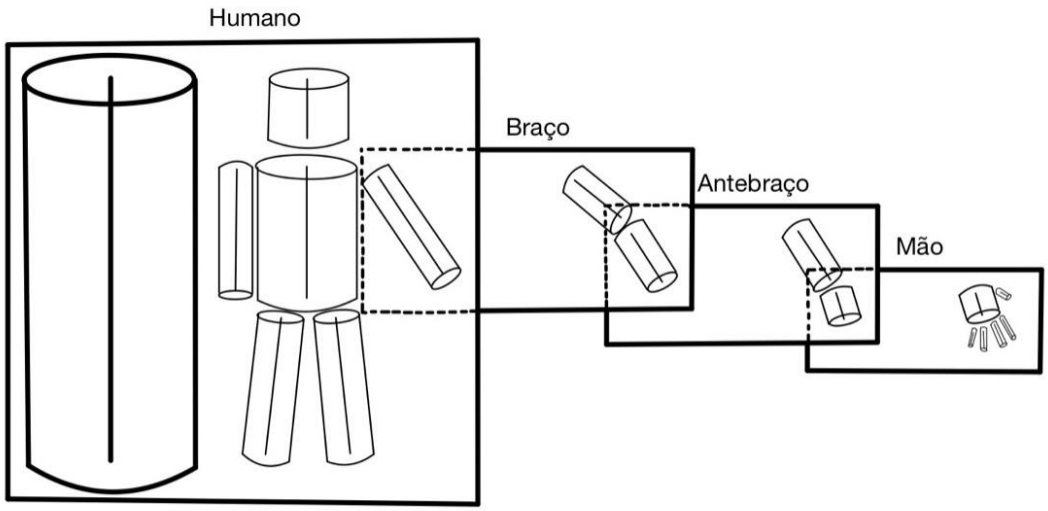

Figura 1.4 - Exemplo de esboço 3D do sistema visual primário (adaptado de Marr (1982)).

Finalmente, a investigação deve prosseguir para o nível implementacional, no qual serão descritos quais subsistemas neurais estão envolvidos no processamento da informação visual e quais processos neurofisiológicos correspondem aos processos algorítmicos descritos nos níveis acima.

O trabalho de David Marr representa um paradigma a respeito de como as ciências cognitivas devem ser conduzidas. Subjacente a esse paradigma está a ideia de uma hierarquia entre as disciplinas. De maneira simplificada, Marr privilegia os níveis mais altos de investigação, os quais informam os níveis hierarquicamente mais baixos o que o cérebro faz e, portanto, o que esses níveis mais baixos devem procurar explicar. Contemporaneamente, muitos autores abandonaram esse paradigma metodológico, reforçando uma maior troca de conhecimento entre os diferentes níveis de abstração do cérebro (ver §4.1 do Capítulo 4). Contudo, o caráter top-down de explicação é, em grande medida, adotado por Fodor. Sua teoria parte da premissa de que a psicologia de senso comum está correta em sua maior parte e tenta, a partir disso, formular hipóteses a respeito dos mecanismos mentais subjacentes aos fenômenos que envolvem as atitudes proposicionais, o raciocínio etc. Para explicar esses fenômenos, Fodor postula a existência de uma linguagem do pensamento.

1.5 A LINGUAGEM DO PENSAMENTO

Um modo de entender a TRM é dizer que ela é uma teoria sobre a *arquitetura da mente*. Ela versa a respeito de

(i) como o cérebro representa as informações que ele processa (dimensão semântica);

(ii) quais tipos de operações ocorrem sobre essas representações (dimensão causal).

O problema (i) deve ser entendido do seguinte modo: em que "formato" o cérebro codifica informação? Existem várias maneiras de representar algo. Um mapa cartográfico é um tipo de representação, o velocímetro de um carro é outro, o sistema binário, outro, e assim por diante. Mas, há uma diferença importante entre dois grandes grupos de representações: as analógicas e as digitais. Representações analógicas (como mapas e velocímetros) utilizam medidas de quantidade proporcionais aquilo que é representado. Se o Congo for maior que a Somália, a representação do Congo será proporcionalmente maior que a representação da Somália; do mesmo modo, o ponteiro de um velocímetro se move proporcionalmente à variação de velocidade real do carro. Já as representações digitais codificam a informação simbolicamente. O fato de o número 1000 ser muito maior do que o número 1 não faz com que a representação binária do número 1000 seja maior que a representação binária do número 1. De acordo com a TRM, a informação é simbolicamente estruturada no cérebro. Há um nível de descrição das atividades cerebrais em que é correto afirmar que estruturas simbólicas com conteúdo informacional estão sendo processadas.

Símbolos são sempre instanciados por meio de uma estrutura física concreta. Chamaremos essa estrutura de *veículo representacional*. Computadores modernos, por exemplo, manipulam dígitos binários. Há uma gama irrestrita de substratos físicos que podem servir de veículo, mas, em geral, usa-se polaridades magnéticas positivas e negativas, ou níveis de tensão elétricos. No caso do cérebro, se ele instancia de fato estruturas simbólicas, os veículos físicos de representação devem ser, presumivelmente, estruturas neurais.

Segundo a TRM, portanto, o que ocorre quando instanciamos uma atitude proposicional é que instanciamos uma estrutura simbólica com um conteúdo proposicional específico, e essa estrutura simbólica cumpre o papel causal associado àquela atitude em nossa economia cognitiva. Se Pedro acredita que Maria irá se casar com Antônio, de acordo com a TRM, Pedro instancia uma estrutura simbólica cujo conteúdo é *Maria irá se casar com Antônio*, e essa estrutura simbólica cumpre as funções associadas ao estado de crença. Se Marcelo duvida que Maria irá se casar com Antônio, então Marcelo instancia uma estrutura simbólica de mesmo tipo, mas ela

cumpre as funções associadas ao estado de dúvida em sua dinâmica psicológica. A metáfora bem conhecida nesse caso é o de que possuímos caixas mentais das atitudes, e podemos guardar símbolos com conteúdo dentro delas. Ter a expectativa de que o dólar irá se desvalorizar com relação ao real é colocar um símbolo com o conteúdo *o dólar irá se desvalorizar com relação ao real* na caixa de expectativas[18].

Mas por que símbolos? Por que não representações de outro tipo, como as analógicas? Como veremos adiante, símbolos possuem a chave para explicar o ponto (ii) acima, a respeito das operações executadas pelo cérebro. Lembre-se que a explicação deve ser sensível ao paralelismo que existe entre, de um lado, as relações causais entre atitudes proposicionais e, de outro, as relações lógicas que existem entre os conteúdos das atitudes. Nosso raciocínio apresenta largamente a valiosa propriedade de *preservar a verdade* de um estado mental para outro, uma propriedade típica de argumentos válidos. Passamos da crença de que a Torre Eiffel fica em Paris & Paris fica na França para a crença de que a Torre Eiffel fica na França; ou da crença de que todo político é corrupto para a crença de que esse político é corrupto. Recorde também que a psicologia de senso comum se compromete com explicações causais. A crença de que a Torre Eiffel fica em Paris & Paris fica na França *causa* a crença de que a Torrei Eiffel fica na França; a crença de que todo político é corrupto *causa* a crença de que esse político é corrupto. O desafio consiste em explicar como essa sequência causal de estados psicológicos é sensível às relações semânticas dos conteúdos que elas expressam.

Essa tarefa é extremamente complicada porque *conteúdo representacional* é uma noção extremamente abstrata, e é intrigante pensar em como o cérebro, que é um sistema físico concreto, pode ser causalmente sensível a ela. Não é difícil imaginar como neurônios engajam em transações causais com outros neurônios. O que é misterioso é como neurônios engajam em transações causais com outros neurônios *em função da informação que eles carregam e das propriedades semânticas que eles possuem* (Bermúdez 2005, p. 73).

1.5.1 O FUNCIONALISMO SOBRE CONTEÚDO REPRESENTACIONAL

Há uma abordagem funcionalista radical que tenta resolver esse problema de maneira direta. Essa teoria propõe que o conteúdo de nossos estados mentais deva ser, assim como a atitude, funcionalmente caracterizado (ver Seção 1.3). A função cognitiva que um estado mental cumpre em nossa dinâmica psicológica seria suficiente para diferenciar não apenas tipos de

[18] "To believe that such and such is to have a mental symbol that means that such and such tokened in your head in a certain way; it's to have such a token 'in your belief box', as I'll sometimes say" (Fodor 1987, p. 17).

estados mentais (e.g. diferenciar crenças de desejos), mas também subtipos de estados mentais (e.g. diferenciar a crença de que P da crença de que Q). A crença de que vai chover e a crença de que a economia vai melhorar são atitudes com conteúdos diferentes porque, no fundo, elas cumprem papéis cognitivos distintos em nosso cérebro. Essa visão é conhecida na literatura como *Conceptual Role Semantics* (CRS)[19], e é uma teoria que promete uma explicação mecanicista e naturalista da cognição[20]. Temos uma vaga ideia a respeito do que seja *conteúdo*, mas não há consenso algum na filosofia sobre o que esse termo designa exatamente. Não seria exagero dizer que esclarecê-lo é um dos problemas mais intratáveis da filosofia da mente, ao lado de problemas como O Problema Difícil da Consciência (Chalmers 2010). A CRS se coloca como uma tentativa cientificamente respeitável de solucioná-lo, associando o conceito de conteúdo a noções que já são bem aceitas pelas ciências naturais. Fica claro como a CRS responde o quebra-cabeça acima. As transições causais entre estados mentais são sensíveis aos seus conteúdos porque as propriedades causais de um estado mental são precisamente o que fixam o seu conteúdo.

Contudo, a CRS sofre muitas dificuldades teóricas (Bermúdez 2005, Fodor 1992). Em primeiro lugar, parece ser extremamente difícil caracterizar, em termos puramente funcionais, qualquer atitude proposicional, por mais simples que seja o conteúdo envolvido. Uma segunda dificuldade, mais grave do ponto de vista teórico, é que essa teoria falha em explicar o que há de comum em situações nas quais dois indivíduos (ou o mesmo indivíduo em momentos distintos) possuem atitudes diferentes frente ao mesmo conteúdo. Digamos que Pedro *acredita* que vai chover e Paula *duvida* que vai chover. Uma análise clássica nos permitiria concluir que os dois indivíduos compartilham o mesmo conteúdo proposicional — instanciam a mesma estrutura simbólica, na TRM —, embora mantenham atitudes diferentes frente a ele. Essa abordagem parece inexplicável para um defensor da CRS, já que o conteúdo, neste caso, é fundamentalmente caracterizado por suas propriedades funcionais. Note que a crença de que irá chover e a dúvida de que irá chover são duas atitudes proposicionais com características psicofuncionais completamente distintas (Pedro e Paula farão inferências diferentes, tomarão decisões diferentes e, como consequência, agirão de maneiras distintas). Se não há nenhuma sobreposição entre as disposições mentais ou comportamentais de quem *acredita que P* e quem *duvida que P* (e, além disso, P é fixado por essas disposições, como quer a CRS), não há sentido em dizer que essas duas pessoas mantêm atitudes diferentes frente à mesma proposição. É

[19] Ver Whiting (Internet Encyclopedia of Philosophy).

[20] Contudo, como Fodor (1985) enfatiza, se a CRS não é uma teoria reducionista – i.e. se ela não elimina o conteúdo – mas apenas defende que conteúdo mental possui propriedades funcionais essencialmente, então ela simplesmente assume que propriedades semânticas estão presentes de alguma forma, sem explicações adicionais. Nesse sentido, ela não seria uma teoria mais naturalista do que a TRM, por exemplo.

duvidoso, inclusive, que a afirmação de que elas discordam com relação a P possua algum significado nesse caso (deve haver algo em comum que é o objeto de discordância, e a CRS desaparece com esse objeto). Parece um preço extremamente alto a se pagar.

Ademais, nossa principal preocupação aqui é a de responder como as transições entre estados mentais respeitam as relações lógicas entre seus conteúdos, e a CRS parece incapaz de fornecer uma explicação para esse fenômeno.

1.5.2 AS PROPRIEDADES SINTÁTICAS DO PENSAMENTO

Considere a proposição *Susan é psicóloga de robôs*. Há um sentido importante no qual o conteúdo expresso por ela é estruturado. Por estruturado quero dizer que ele é 'quebrável' em partes menores, e que essas partes possuem um arranjo específico. Tecnicamente, o conteúdo proposicional é uma função 1) do conteúdo dos termos (conceitos) menores que o constituem (*Susan, é, psicóloga de robôs*) e 2) do modo como esses termos (conceitos) se relacionam. De modo que se trocássemos algum dos termos ou invertêssemos sua ordem, teríamos uma proposição diferente (Szabó 2020). Considere, agora, a *sentença* "Susan é psicóloga de robôs". Existe um isomorfismo estrutural entre as duas, pois para cada elemento da sentença existe um elemento da proposição e, além disso, as relações entre os elementos da proposição são perfeitamente transportadas para os elementos da sentença[21]. Podemos falar de estrutura tanto no nível semântico quanto no nível do veículo representacional. Mas, note que não é uma regra que o veículo representacional seja sempre isomórfico à proposição que ele expressa. Poderíamos, por exemplo, convencionar um símbolo qualquer e dizer que esse símbolo expressa o conteúdo *Susan é psicóloga de robôs*.

A Hipótese da Linguagem do Pensamento (Fodor 1975) afirma que os veículos do pensamento são análogos às sentenças línguas naturais, no sentido de que as proposições que são objetos de nossas atitudes proposicionais são instanciadas por estruturas isomórficas a elas, como ocorre com as sentenças do português, inglês etc[22]. Assim, se Tom acredita que Susan é psicóloga de robôs, de acordo com a HLP, Tom instancia uma estrutura simbólica (uma sentença do *mentalês*) isomórfica à proposição *Susan é psicóloga de robôs*.

---

[21] Isomorfismo estrutural é uma relação caracterizada por (i) uma função f bijetora entre duas estruturas $\langle A, R\rangle$ e $\langle B, S\rangle$, tal que para cada elemento $a \in A$ há um elemento correspondente $b \in B$ e (ii) para qualquer relação entre os elementos de uma das estrutura, R(x, y) se, e somente se, $S(f(x), f(y))$ (Enderton 1977). Um exemplo familiar é a relação de isomorfismo entre um mapa geográfico e o território representado.

[22] "Some philosophers apparently think it's an insightful argument against this sort of theory to hoist their eyebrows and say: 'You aren't *really* supposing that there are *sentences* in the *head*.' 'Yes I am' is the equally insightful reply". Fodor (2001, p. 107).

Aqui reside a diferença fundamental entre a Teoria Representacional da Mente de Fodor e a Conceptual Role Semantics. Embora a CRS possa reconhecer o nível de estruturação semântico de nossas atitudes proposicionais, ela não pode esperar nenhum tipo de isomorfismo entre esse nível semântico e o nível dos veículos representacionais que instanciam as atitudes. De acordo com a CRS, tanto a atitude quanto o seu conteúdo proposicional são individuados pelo papel funcional que eles exercem no sistema. Não há, em nenhum sentido relevante, estrutura no nível dos veículos físicos do pensamento. Quais recursos teóricos a CRS possui para explicar que a crença de que Susan é psicóloga de robôs tem grande probabilidade de gerar a crença de que alguém é psicóloga de robôs, ou a crença de que alguém é psicóloga? Parece que ela precisa assumir como um fato bruto de nossa psicologia que as propriedades causais de nossos pensamentos mapeiam as relações lógicas entre seus conteúdos[23].

A TRM tenta resolver esse problema combinando a Hipótese da Linguagem do Pensamento com insights de Alan Turing e da lógica formal, compondo a Teoria Computacional da Mente (TCM)[24]. Vejamos como.

Todo conceito (símbolo) possui qualidades formais, as quais definem inteiramente o modo como ele pode se conectar e se relacionar com outros conceitos. No caso da linguagem do pensamento,

(Γ) *as regras de combinação e transformação são realizadas fisicamente por meio das propriedades físicas do veículo representacional que constitui o símbolo.*

Vimos que instanciar uma atitude proposicional, e.g. crença de que João ama Maria, é instanciar uma estrutura simbólica que expressa o conteúdo *João ama Maria*. Vimos também que a estrutura simbólica será isomórfica a esse conteúdo. Portanto, a crença de que João ama Maria envolve instanciar um símbolo JOÃO, um símbolo AMA e um símbolo MARIA, organizados na forma JOÃO AMA MARIA[25]. Dizer que cada um desses símbolos possui propriedades formais é, dentre outras coisas, dizer que eles seguem regras que restringem o modo como

[23] Um outro problema seria o seguinte. Por que alguns estados mentais não possuem conteúdo proposicional, já que eles também cumprem algum papel funcional em nossa psicologia? Os contra-exemplos seriam dor, prazer, ansiedade etc. Esse problema é, na verdade, resultado de um problema mais fundamental. Tratar as atitudes proposicionais como entidades inteiramente funcionais gera a incapacidade de dizer quais propriedades causais fixam a atitude e quais fixam o conteúdo proposicional.

[24] A TCM não é uma teoria, mas um conjunto de teorias sobre a mente inspiradas nas descobertas de Alan Turing. Ver Rescorla (2020). Neste trabalho, o termo *TCM* designa sempre a vertente teórica defendida por Fodor, exceto quando for afirmado o contrário.

[25] Isto é, organizados de tal forma que o conteúdo resultante da concatenação seja *João ama Maria*.

podem ser organizados. Por exemplo, eles também poderiam ser organizados na forma MARIA AMA JOÃO, mas não nas formas MARIA JOÃO AMA e JOÃO MARIA AMA. Mas como essas regras são seguidas? De acordo com **Γ**, os veículos representacionais que constituem os símbolos possuem propriedades físicas tais que garantem a validade dessas regras, mais ou menos do mesmo modo que o formato de uma chave determina quais cadeados ela abre (Fodor 1987, p. 18)[26]. É um corolário de **Γ** que, em última análise, haveria um impedimento *físico* de conectar esses conceitos dessa forma.

Note também que a afirmação **Γ** é neutra com relação ao tipo de material físico que constitui o veículo simbólico. Isso é proposital. A sintaxe pertence a um nível abstrato de organização, um nível funcional. Assim, **Γ** deixa em aberto a possibilidade de dois sistemas de naturezas físicas distintas implementarem o mesmo tipo de sistema simbólico.

Mas isso ainda não explica o paralelismo que existe entre as propriedades causais das atitudes proposicionais e as relações lógicas entre os seus conteúdos. O que explica isso, segundo a TRM, é o fato de que

(**Ψ**) *as propriedades sintáticas da linguagem do pensamento espelham suas propriedades semânticas.*

Isso garante que as sentenças formadas possuam sentido (e sequências como JOÃO MARIA AMA sejam impossíveis de se formar), mas, além disso, as regras formais asseguram que as transições entre proposições simbólicas respeitem as relações lógicas entre os seus conteúdos. O insight teórico sintetizado em **Ψ** é fruto dos avanços recentes da lógica formal (Mortari 2001, Haack 1998). A disciplina de lógica, como a entendemos hoje, contrasta de maneira muito evidente a dimensão sintática da dimensão semântica da linguagem. Isso significa que podemos entender sentenças lógicas de uma perspectiva puramente formal ou, se quisermos, de uma perspectiva semântica. Podemos, além disso, analisar as relações que existem entre esses dois níveis. Com efeito, é uma característica de certos sistemas formais que as propriedades sintáticas dos símbolos "imitam" suas propriedades semânticas. Considere a lógica de primeira ordem.

[26] A analogia com chaves e cadeados é interessante porque, assim como não é necessário que exista uma regra intencional que dite como chaves e cadeados se conectam, não é necessário que existam regras intencionais que ditem como os símbolos se conectam. Nós podemos descrever o modo como eles se conectam a partir de regras gerais, mas elas não precisam ser representadas dentro do sistema. "The rules which determine the course of such transformations may, but needn't, be themselves explicitly represented. But the mental contents (the 'thoughts', as it were) that get transformed *must* be explicitly presented or the theory is simply false" (Fodor 1992, p. 290).

A lógica de primeira ordem (chamemos £) possui uma série de símbolos e regras de manipulação simbólica. O alfabeto simbólico é dividido em categorias sintáticas. Reservamos as letras maiúsculas para relações, propriedades e proposições, e as letras minúsculas para indivíduos e variáveis. Também temos os conectivos verofuncionais (cuja função é concatenar proposições) e os quantificadores (que, no caso da lógica de primeira ordem, quantificam sobre indivíduos). Assim, a respeito da sintaxe, é uma regra em £ que, se quisermos representar que o indivíduo *a* possui a propriedade *P*, escrevemos *P* seguido de *a*: *Pa*. Do ponto de vista sintático, porém, podemos esquecer que a fórmula *Pa* expressa a proposição de que um indivíduo *a* possui a propriedade *P*. Podemos simplesmente atentar para as características formais dos símbolos, viz., que é uma letra maiúscula seguida de uma letra minúscula. De acordo com as regras sintáticas de £, *Pa* é uma *fórmula bem formada*, e esse fato depende unicamente das regras de composição, independentemente do significado que atribuímos aos símbolos. De acordo com as regras sintáticas, por exemplo, sequências como *AA* ou *bb* não são fórmulas bem formadas, são fórmulas sintaticamente mal formuladas. Isso não é sem razão. As regras sintáticas foram cuidadosamente pensadas para que, a partir delas, apenas fórmulas com conteúdo semântico claro possam ser formadas, e *AA* e *bb* seriam semanticamente ininteligíveis (assumindo que maiúsculas expressam relações, propriedades e proposições e minúsculas expressam indivíduos). Nesse sentido, as regras sintáticas respeitam as relações semânticas dos símbolos no que diz respeito às suas *combinações possíveis*: a sintaxe garante que aquilo que tenha sentido possa ser composto, e aquilo que não tenha sentido, não. Esse é um primeiro nível em que a sintaxe espelha a semântica.

As regras sintáticas de um sistema formal também espelham a dimensão semântica num nível mais profundo: a respeito de como *os conteúdos das proposições simbólicas estão inferencialmente conectados*. Por exemplo, é uma regra de dedução da lógica de primeira ordem[27] que se temos $A \rightarrow B$ numa linha e A noutra linha, podemos escrever B. O que significa que B é uma *consequência sintática* de $A \rightarrow B$, A em £, ou, em outras palavras, o conjunto de premissas formado por $A \rightarrow B$, A *deduz formalmente* B. Na notação lógica, escrevemos $A \rightarrow B$, $A \vdash_{£} B$.

1.	$A \rightarrow B$	Premissa
2.	A	Premissa
3.	B	1, 3 *Modus Ponens*

[27] Mortari (2001, cap. 14). O autor apresenta um sistema de dedução natural não axiomático.

Chamamos as três linhas acima de uma *derivação em £*, que consiste na manipulação de símbolos de acordo com suas formas tipográficas e regras sintáticas de £, sem se ocupar do conteúdo ou significado dos símbolos. A forma argumentativa acima é tão comum que recebe um nome, Modus Ponens. Mas note que, embora essa derivação seja 'cega' para o conteúdo das sentenças simbólicas que estão sendo manipuladas, a regra é concebida conforme nossa intuição de que se as premissas forem verdadeiras, a conclusão também terá de ser. Considere o argumento a seguir, apresentado em língua natural, com a mesma estrutura Modus Ponens.

P1. Se Sócrates for ser humano, então Sócrates veio de Júpiter.
P2. Sócrates é ser humano.
C1. Sócrates veio de Júpiter.

Uma das premissas é falsa, e a conclusão também o é. No entanto, C1 é uma consequência semântica de P1 e P2 (P1, P2 ⊨ C1), porque se fosse o caso que as premissas fossem verdadeiras, a conclusão também teria de ser[28]. Note que quando falamos de implicação semântica, não estamos preocupados com regras sintáticas e formas tipográficas. Em vez disso, fazemos referência a noções como verdade e falsidade. Formalmente, dizemos que α (uma proposição qualquer) é uma consequência semântica de Δ (um conjunto não-vazio de premissas) sempre que a verdade das premissas de Δ implique na verdade da conclusão α. Na notação lógica, escrevemos $\Delta \vDash \alpha$[29].

Um argumento pode ser analisado por nesses dois níveis diferentes, um nível sintático (que leva em consideração as características tipográficas dos símbolos e as regras inferenciais do sistema) e um nível semântico (que leva em consideração a verdade e falsidade das premissas e conclusão). O ponto a ser destacado é o de que a regra sintática Modus Ponens mapeia as relações lógicas entre premissas e conclusão (não obstante a regra, ela mesma, seja 'cega' a respeito dessas relações). Outro modo de dizer isso é dizer que a regra *preserva a verdade*; i.e., se as premissas forem verdadeiras, a conclusão também será (a verdade é sistematicamente 'carregada' das premissas para a conclusão). No caso da lógica proposicional, graças ao trabalho

[28] Laird (1980) argumenta que, do ponto de vista psicológico, tentamos construir modelos que falsifiquem o argumento (i.e. modelos que tornem as premissas verdadeiras e conclusão falsa). Se não conseguirmos construir tal modelo, tomamos o argumento como válido. Ver também Haack (1998).

[29] Não introduzo a noção de modelos por acreditar que o assunto se tornaria mais complicado do que o necessário (levando em conta o objetivo dessa exposição). Sacrifico, assim, um pouco do rigor lógico em troca de mais simplicidade para o texto.

de Kurt Gödel (1986), podemos extrapolar essa propriedade para todas as regras de inferência do sistema. Todas as regras preservam as relações lógicas entre premissas e conclusão, o que significa que o sistema é *correto* (*sound*). Gödel, na verdade, prova algo muito mais forte do que a correção da lógica proposicional. Em 1929, o matemático também prova que o sistema é *completo*, o que significa que todas as verdades expressáveis pelo vocabulário do sistema são demonstráveis pelas regras de inferência. Ou seja, se um conjunto de premissas implica semanticamente uma proposição, então haverá uma derivação sintática das premissas para a conclusão.

(Teorema da Completude e Correção): $\Delta \vDash \alpha$ *se, e somente se,* $\Delta \vdash \alpha$.

Gödel prova que, na lógica proposicional, a noção sintática e a noção semântica coincidem. A lógica proposicional possui a propriedade de *correção*, o que significa que as regras sintáticas nunca derivam falsidades (i.e. aplicando as regras de manipulação simbólica, de uma verdade nunca se seguirá uma falsidade); e a lógica proposicional é *completa*, o que significa que todo argumento semanticamente válido expressável na linguagem simbólica possui uma demonstração sintática (ou seja, a completude garante que as regras de inferência dão conta de deduzir a conclusão a partir das premissas, caso a conclusão seja uma consequência lógica das premissas). As regras sintáticas da lógica proposicional imitam de maneira perfeita as relações semânticas entre os conteúdos dos símbolos.

O Teorema da Completude e Correção de Gödel ilustra a possibilidade teórica de existirem sistemas que exibem o tipo de relação especificado em **Ψ** acima. Essa ideia é central na Teoria Computacional da Mente, pois nos permite entender de que modo as transições entre estados mentais podem ocorrer mecanicamente, como o resultado de interações causais "cegas" entre estruturas físicas (simbólicas), concomitante ao fato de que essas interações preservam as relações semânticas entre os conteúdos dessas estruturas.

Assim, a TCM afirma que:

As representações internas da linguagem do pensamento são causalmente eficazes em função de suas propriedades semânticas via suas propriedades sintáticas. Todas as transições entre estados representacionais ocorrem em virtude das propriedades formais dos símbolos da linguagem do pensamento. Como afirma **Γ**, essas propriedades sintáticas são fisicamente realizadas pelas propriedades materiais do veículo simbólico. Em última análise, portanto, o raciocínio consiste em interações causais no nível neural. Processos mentais são processos físicos. Todavia, esses processos neurais instanciam um padrão sintático, que diz respeito a como

símbolos com conteúdo semântico se conectam e se relacionam. O encadeamento lógico entre estados mentais também ocorre via sintaxe, pois as propriedades sintáticas da linguagem espelham de maneira adequada as propriedades semânticas dos símbolos. Nossos processos racionais são, no fundo, sensíveis apenas às propriedades causais de estruturas neurais. Ocorre que essas propriedades causais instanciam regras sintáticas (formais) que, por sua vez, coincidem com os atributos semânticos dessas estruturas neurais, de maneira análoga ao que ocorre com a lógica proposicional e outros sistemas formais[30].

John Haugeland batiza os sistemas que manipulam símbolos de maneira cega — embora respeitando as restrições semânticas dos símbolos — de *semantic engines*, ou máquinas semânticas[31]. O paradigma das máquinas semânticas é o computador. Alan Turing mostrou que é possível criar máquinas que são sensíveis às propriedades formais de uma proposição, precisamente porque essas propriedades formais podem ser reduzidas às propriedades físicas do veículo que a codifica. Em outras palavras, Turing mostrou como construir máquinas capazes de detectar e produzir inferências válidas, porque a validade de uma inferência diz respeito unicamente à sua forma. Como as transições racionais entre estados mentais instanciam a mesma relação que existe entre premissas e conclusões de um argumento válido, Turing possivelmente nos revelou a *mecânica* da cognição.

> The trick is to combine the postulation of mental representations with the 'computer metaphor'. Computers show us how to connect semantical with causal properties for symbols. So, if having a propositional attitude involves tokening a symbol, then we can get some leverage on connecting semantical properties with causal ones for thoughts. (FODOR, 1988, p. 18).

1.6 A ANALOGIA COMPUTACIONAL

A Teoria Computacional da Mente é conhecida por endossar slogans como "o cérebro é um computador" ou "a mente é o software do cérebro". Acredito que algumas ressalvas precisam ser feitas, embora a ideia geral esteja correta. Pode ser instrutivo pensar na relação entre mente e cérebro do mesmo modo que pensamos na relação entre o software e o hardware de um computador. O primeiro problema, porém, é que não é claro que tipo de relação é essa. A TCM, por exemplo, é uma teoria consistente tanto com correntes filosóficas dualistas quanto com correntes filosóficas materialistas. Já que o mesmo software (ou estado psicológico) pode ser

[30] A completude é uma propriedade muito especial de sistemas formais. Nem todos os sistemas formais são completos. Gödel prova que a aritmética é incompleta: existem verdades na aritmética que não podem ser demonstradas. Para uma introdução extensa a respeito dos teoremas de Gödel (e algumas ponderações sobre como isso se relaciona com a mente) ver Smith (2013).

[31] Encontrei esse termo em Heil (2013), o qual faz referência a John Haugeland.

implementado por sistemas físicos completamente distintos, o dualista argumenta que ele é algo distinto da matéria ou do sistema material que o implementa; o dualista pode argumentar a favor de um dualismo de substância ou de um dualismo de propriedade, i.e., ele pode argumentar que existem dois tipos diferentes de substâncias (à la Descartes, res extensa e res cogita), ou dois tipos diferentes de propriedades (propriedades físicas e propriedades mentais). O materialista, por outro lado, argumenta que o software (ou estado psicológico) é apenas uma *organização* da matéria, mas que não há nenhum tipo de substância ou propriedade diferente da material, apenas abstrações que fazemos a respeito da estrutura causal do sistema e de como suas partes estão arranjadas.

Fodor, do modo como eu o entendo, parece defender uma vertente materialista. Em seu livro *Psychological Explanation*, de 1968, ele tece uma série de críticas à corrente behaviorista radical que o antecede, a qual ele considera absurda a ponto de ser difícil de explicar como ela pôde ter adquirido tantos adeptos. O behaviorismo radical é a tese de que todo predicado mental (e.g. possuir a crença de que P) pode ser literalmente definido em termos comportamentais, de modo que há uma *conexão lógica* entre predicados psicológicos e predicados comportamentais (Fodor 1968, p. 50-51). Em última análise, falar de estados mentais é falar de disposições comportamentais. (Esse tipo de behaviorismo – o qual já não encontra mais tantos defensores – fincava suas raízes no positivismo lógico e, portanto, esforçava-se para que seus termos designassem eventos e entidades observáveis, a fim de respeitar a tese *verificacionista* que caracterizava o positivismo)[32]. Além das críticas, Fodor oferece um diagnóstico para explicar por que o behaviorismo chegou a ser a teoria dominante de sua época. De acordo com o autor, acreditava-se que o behaviorismo era a única opção teórica ao dualismo cartesiano e que, portanto, negar aquele era o mesmo que endossar este último[33]. Sua saída para este engano é definir um termo alternativo às duas correntes, o *mentalismo*[34]. O mentalismo é definido por Fodor como a negação da tese behaviorista de que há uma conexão lógica entre predicados psicológicos e predicados comportamentais, de modo que mentalismo e behaviorismo são termos exaustivamente excludentes. Além disso, o mentalismo é uma opção teórica ao dualismo,

[32] "Don't human beings talk of introspectible entities, thoughts, feelings, and so on, even if these are not recognized by behaviorism or best understood as behavioral tendencies? Psychological behaviorists regard the practice of talking about one's own states of mind, and of introspectively reporting those states, as potentially useful data in psychological experiments, but as not presupposing the metaphysical subjectivity or non-physical presence of those states. There are different sorts of causes behind introspective reports, and psychological behaviorists take these and other elements of introspection to be amenable to behavioral analysis" (Graham, 2019).

[33] "We have seen that one possible way of arguing against dualism is to establish some form of behaviorism, and that, in point of historical fact, a major motivation for adopting behaviorism has been the mistaken supposition that it provides the only viable line of argument against dualism" (Fodor 1968, p. 60). Ver também Fodor 1981.

[34] Fodor não inventou este termo. Ele é usado com significados diferentes por autores diferentes. Ver, por exemplo, Penrose (1994, p. 16) para um uso diferente do termo.

embora seja compatível com ele. O dualismo também defende que predicados psicológicos são logicamente independentes de predicados comportamentais, pelo simples fato de que esses predicados se referem a *tipos* de substâncias diferentes. O mentalista, entretanto, não precisa se comprometer com essa última tese. Ele pode aceitar que, embora predicados psicológicos não possuam nenhuma relação lógica com predicados comportamentais, ambos se aplicam ao mesmo tipo de substância. O mentalismo, portanto, é compatível também com o fisicalismo. Fodor é um mentalista fisicalista, do modo como eu o entendo.

O fisicalismo é a tese de que tudo no universo é físico[35]. Um mentalista fisicalista, portanto, defende que eventos mentais são eventos físicos. Sendo assim, alguém poderia sugerir que processos psicológicos poderiam ser compreendidos e explicados em termos físicos. Fodor argumenta, porém, que o materialismo não implica um reducionismo fisicalista. Ele defende que todo evento mental é um evento físico, mas que a psicologia não pode ser reduzida às ciências naturais, como à física ou à química. O reducionismo clássico exige que, para que uma teoria T_i seja redutível a uma teoria T mais fundamental, é necessário que (i) exista uma tradução entre os vocabulários das duas teorias (*bridge-laws*) que garanta que elas são comensuráveis (i.e. que garanta que elas falem sobre os mesmos fenômenos) e (ii) as leis e princípios de T_i sejam deriváveis a partir das leis e princípios de T (Riel & Gulick 2019). O exemplo paradigmático de redução teórica ocorre entre a termodinâmica e a mecânica estatística (Idem). A termodinâmica descreve o comportamento macroscópico de gases a partir de predicados como temperatura, pressão e volume; a mecânica estatística descreve os mesmos fenômenos de uma perspectiva microscópica, tratando gases como coleções de pequenas moléculas que se agitam de maneira mais ou menos independente, utilizando conceitos como o de energia cinética média das moléculas e distribuições de probabilidade. É possível traduzir os termos de uma teoria para outra – e.g. a noção de temperatura é equivalente à noção de energia cinética média – e, além disso, é possível derivar e explicar os princípios da termodinâmica a partir da mecânica estatística[36].

Mas o problema de reduzir a psicologia às ciências mais básicas é que há a possibilidade empírica de que um tipo de evento individuado por um predicado psicológico corresponda a uma disjunção heterogênea de eventos na ciência mais básica. Assim, um princípio $\mathcal{P}$ da psicologia que relacione dois predicados mentais, como

[35] Ver Stoljar (2015) para uma exposição introdutória do tema.

[36] Na verdade, essa é uma simplificação grosseira. Nem todos os termos teóricos da termodinâmica são perfeitamente traduzidos para a mecânica estatística, e nem todos as leis da termodinâmica são deriváveis a partir das leis da mecânica estatística. Ver Sklar (1993).

$\mathcal{P}$: P → D

digamos, dor gera comportamento aversivo, seria equivalente, na ciência mais básica, a um princípio que relacione dois grupos de eventos heterogêneos entre si (i.e. heterogêneos com relação aos elementos do próprio grupo), o que parece absurdo. Se estados mentais forem funcionalmente caracterizados, um predicado psicológico como dor corresponderá a uma disjunção (não exclusiva) assistemática de predicados neurais que se aplicam a humanos e possivelmente a muitas outras espécies[37]; a redução clássica é inviável porque um predicado mental P que individua um tipo de evento na psicologia não corresponde a um tipo de evento na ciência mais básica.

Redução $\mathcal{R}$ do princípio $\mathcal{P}$:

$\mathcal{R}$: $(P_1 \vee P_2 \vee ... \vee P_n) \rightarrow (D_1 \vee D_2 \vee \ldots \vee D_k)$, onde $1 \leq n \leq p$, e $1 \leq k \leq q$.

O argumento de Fodor é o de que $\mathcal{R}$ não pode ser considerada uma lei, porque nem o antecedente e nem o consequente constituem *tipos* naturais. Ele ilustra esse ponto do seguinte modo:

> I think, for example, that it is a law that the irradiation of green plants by sunlight causes carbohydrate synthesis, and I think that it is a law that friction causes heat, but I do not think that it is a law that (either the irradiation of green plants by sunlight or friction) causes (either carbohydrate synthesis or heat) (FODOR 1975, p. 21).

Note também que há a possibilidade de os dois conjuntos de disjuntos de $\mathcal{R}$ não terem a mesma cardinalidade, e que é possível que algum Pi não tenha um Di correspondente. Isso explicaria porque as ciências especiais como a psicologia e a economia possuem exceções, e seus princípios são apenas legiformes (em oposição aos princípios da física e da química, que constituem leis que não admitem exceções). É possível que o antecedente de $\mathcal{R}$ seja satisfeito, mas que não haja um consequente correspondente – é possível imaginar casos em que haja dor sem comportamento aversivo –, e tudo isso é consistente com a ideia de que as leis das ciências mais básicas são invioláveis[38].

O ponto de Fodor é que todo evento mental possui uma *descrição* física. Mas o que faz com que um evento possua uma natureza psicológica (ou econômica, ou biológica etc.) é seu enquadramento nas generalizações da psicologia (ou da economia, ou da biologia etc.). Fodor insiste que esse não é um fato epistêmico acerca de como nós conhecemos ou compreendemos o mundo. Ele afirma que

---

[37] Digamos, fibras-C em humanos, mas disparo do neurônio #789 em cachorros.

[38] Fodor desenvolve melhor todos esses pontos em sua introdução de The Language of Thought, de 1975.

> [...] there are special sciences not because of the nature of our epistemic relation to the world, but because of the way the world is put together: not all the kinds (not all the classes of things and events about which there are important, counterfactual supporting generalizations to make) are, or correspond to, physical kinds (FODOR, 1975, p. 24).

Nesse sentido, a analogia entre a mente e o software de um computador é certeira: os dois pertencem a um nível de descrição do sistema que se abstrai de sua constituição e implementação física, de modo que esse nível de descrição é irredutível ao nível de descrição físico.

Em contrapartida, porém, o cérebro e o hardware de um computador são sistemas muito diferentes, a começar pelos seus componentes. Os neurônios, as principais células do nosso aparato cognitivo, são estruturas que se conectam com centenas de seus vizinhos, compondo malhas extremamente complexas, enquanto que os componentes de um computador fazem menos de 10 conexões adjacentes. Componentes eletrônicos são interruptores biestáveis, i.e., eles possuem dois estados: ligado ou desligado, 1 ou 0. Neurônios, por outro lado, não podem ser caracterizados dessa forma, já que mesmo quando eles não estão disparando sinais elétricos, essas células realizam atividades cognitivamente relevantes (Bermúdez 2014, p. 108-109). Computadores executam tarefas serialmente, e há boas razões para acreditar que o cérebro implementa algum tipo de processamento paralelo[39]. O armazenamento e acesso à memória também parece diferir absolutamente (Copeland 1993). E a lista continua.

No entanto, essas diferenças são superficiais em certo sentido. Elas escondem uma semelhança muito mais profunda a respeito de como os dois sistemas operam. Ambos são essencialmente manipuladores de informação (*informavores*, como Pylyshyn (1984) batiza). Seus estados internos representam, de alguma forma, objetos externos, eventos, ideias abstratas, coordenadas espaciais e muitas outras coisas. Esse conteúdo é, de algum modo, manipulado para que o sistema possa executar tarefas de rotina, resolver problemas, planejar ações, prever consequências, tomar decisões etc. O computador resolve esses problemas por meio da implementação de uma *linguagem interna*. O conteúdo de seus estados internos é codificado por meio de *símbolos*; e todas as operações implementadas por ele são realizadas através dos veículos

[39] Neste contexto, frequentemente menciona-se a regra dos 100 passos. Considera-se que o tempo que o cérebro demoraria para realizar um passo computacional é de 5ms, baseado no tempo que um neurônio leva para gerar um potencial de ação (enviar um sinal elétrico). Muitas tarefas cognitivas, porém, levam um curtíssimo período de tempo para serem realizadas. Reconhecimento visual, por exemplo, leva em torno de 500ms. Se o cérebro operasse serialmente, isso significaria que o algoritmo que ele utiliza para reconhecer objetos precisaria de apenas 100 passos. Isso é extremamente improvável, dada a complexidade da tarefa. Ver Copeland (1993, p. 281) e Bermúdez (2005, p. 107).

que constituem sua *linguagem de máquina*. Turing (1936) revelou que é possível manipular mecanicamente informações respeitando as propriedades semânticas que elas carregam, pois o veículo simbólico que encripta a informação instancia propriedades sintáticas que refletem suas distinções semânticas relevantes. A hipótese computacionalista é a de que o cérebro também opera por meio da implementação de uma *linguagem interna*. Neurônios codificam conteúdos informacionais através de *símbolos*, e as operações que ele implementa são realizadas através dos veículos que constituem uma *linguagem do pensamento*.

É uma imprecisão, portanto, dizer que o cérebro é um computador, pois os dois trabalham de maneiras muito diferentes na execução de suas tarefas. Em essência, porém, eles são o mesmo tipo de sistema: sistemas computacionais. São máquinas que processam a informação através de uma linguagem simbólica própria, unicamente por meio das propriedades formais dos símbolos que elas manipulam.

Mas o que significa dizer que o cérebro instancia um sistema linguístico interno? Evidentemente, se abrirmos o cérebro não veremos símbolos inscritos na massa encefálica, do mesmo modo que não vemos 0's e 1's transitando nos circuitos de um computador. A linguagem pertence a um nível de abstração neutro com relação aos detalhes físicos da implementação. A ideia é que as representações internas postuladas pelo nível de descrição psicológico – as quais servem de matéria prima dos processos cognitivos – são implementadas fisicamente por meio de veículos estruturados. Esses veículos são linguísticos porque existe um isomorfismo estrutural entre eles e as proposições que eles expressam. Isso é fácil de entender quando pensamos no computador. Seus estados internos representam 0's e 1's em virtude desse isomorfismo, i.e., em virtude da existência de tipos de estados internos que correspondem a 0's e tipos de estados internos que correspondem a 1's; esses estados internos podem ser realizados, por exemplo, por campos magnéticos (campo magnético positivo = 1 e campo magnético negativo = 0), por presença ou ausência de corrente, por válvulas de água que abrem e fecham; por um conjunto irrestrito de objetos. O que é realmente importante é que existam estados internos que possam ser interpretados como 0's e 1's e que os veículos físicos que os realizam tenham propriedades causais que espelhem as relações semânticas dos conteúdos que eles expressam.

Mas há uma diferença. Os estados internos de um computador representam 0's e 1's porque nós *atribuímos* esses valores a eles. Que o campo magnético positivo representa o valor 1 e o campo magnético negativo representa o valor 0 é, em certo sentido, uma propriedade relativa a nós, observadores. Tanto é esse o caso que poderíamos inverter a interpretação, e fazer com que campos positivos representassem 0 e campos negativos representassem 1. Esse ponto

parece se generalizar. Tudo o que consideramos símbolos carregam significado apenas em relação a nós ou a algum observador. *Ser um símbolo* parece ser uma propriedade relativa a um observador. Mas se esse for o caso, toda explicação a respeito dos processos cognitivos fundamentada em computações sobre estruturas simbólicas estaria comprometida. Esse problema é levantado por John Searle (1992), um influente crítico da Teoria Computacional da Mente. Exporei sua crítica detalhadamente no próximo capítulo.

Por fim, é importante notar o caráter esquemático da Teoria Computacional da Mente. A teoria é silenciosa sobre os tipos de símbolos que o cérebro utiliza (se serão 0's e 1's como os computadores, ou algum outro conjunto de símbolos primitivos), sobre qual nível fisiológico é responsável pela implementação simbólica (i.e., se os neurônios são os blocos de construção da linguagem ou se são estruturas ainda menores), sobre quais faculdades cognitivas são de fato computacionais, e assim por diante. Sem uma hipótese mais detalhada a respeito desses problemas, a teoria se afasta da possibilidade de confirmação ou refutação empírica, e grande parte da discussão a respeito de sua plausibilidade ocorre em domínio filosófico. Esse, no entanto, é o curso natural de todas as ciências.

2. CAPÍTULO 2: CRÍTICA À RAZÃO COGNITIVA

No capítulo anterior vimos que, de acordo com a Teoria Computacional da Mente (TCM) advogada por Fodor (1975; 1987; 1992), o cérebro implementa uma linguagem simbólica por meio da qual a informação é veiculada e processada. Apresentamos a TCM como um fundamento da psicologia de crenças e desejos, ou seja, como uma teoria que fundamentalmente explica as atitudes proposicionais. A ideia é que instanciar a crença de que P envolve instanciar fisicamente uma estrutura simbólica que expressa o conteúdo P e, além disso, essa estrutura física cumpre no sistema o papel funcional associado ao estado mental de crença. Analogamente, instanciar o desejo de que P envolve instanciar fisicamente uma estrutura simbólica que expressa o conteúdo P e, além disso, essa estrutura física cumpre no sistema o papel funcional associado ao estado mental de desejo. O grande mérito da teoria, de acordo com Fodor, é o de explicar mecanicamente como as transições entre estados mentais respeitam as relações lógicas entre os seus conteúdos. Ou como passamos de crenças verdadeiras para crenças verdadeiras. Ou, ainda, como ocorre o raciocínio. De acordo com o autor, o raciocínio consiste na manipulação simbólica através das propriedades sintáticas dos símbolos, as quais mapeiam suas propriedades semânticas, semelhantemente ao que ocorre na lógica formal contemporânea. Turing nos mostrou que é possível construir máquinas que são sensíveis às propriedades sintáticas dos símbolos e, embora totalmente cegas para os conteúdos expressos por eles, operam de forma a respeitar as relações lógicas entre tokens simbólicos. Turing nos revelou como processos computacionais abstratos podem ser implementados fisicamente. A TCM assere que o cérebro é um tipo de sistema computacional, o qual implementa processos computacionais através de estruturas neurais.

Duas ressalvas são bem-vindas aqui. Primeiramente, a teoria não tenta explicar como estados cerebrais possuem conteúdo, ela simplesmente assume que esse é o caso. Ela não tenta explicar como possuímos representações internas acerca de objetos do nosso ambiente, das relações espaciais que existem entre eles, de relações causais entre eventos, de números ou de entidades teóricas microscópicas. Fodor (1975) argumenta que nossas melhores teorias assumem que essas representações internas existem, e que é um princípio metodológico razoável aceitar a ontologia de nossas melhores ciências. Em segundo lugar, a TCM não busca revelar como possuímos estados conscientes e qualitativos. Ela é silenciosa a respeito de quais mecanismos biológicos seriam responsáveis por causar as sensações qualitativas de colocar algo salgado na boca, ou de ouvir o barulho da chuva ou a sensação de satisfação de tirar um sapato apertado depois de um dia de trabalho. É verdade que nossa vida mental é extraordinariamente rica e

complexa em diferentes aspectos e níveis. A TCM, no entanto, discorre sobre apenas um desses importantes aspectos: a inteligência. Quando Alan Turing, em seu artigo de 1950, dispensou a pergunta "Podem as máquinas pensar?" como uma questão mal formulada, era porque não havia nenhuma definição do termo *pensar*. Se não há definição bem aceita do que signifique esse termo, a questão de se as máquinas podem pensar ou não se torna simplesmente terminológica. Seria como perguntar se os submarinos nadam: depende do que se entende pelo termo *nadar*. A TCM corresponde a uma tentativa de revelar a natureza oculta do termo *pensar*, a definição que precisamos para responder à pergunta que Turing outrora rejeitou. A ideia fundamental é que o raciocínio está profundamente relacionado com a atividade de manipular e transformar símbolos[40].

Nem todos concordam com essa abordagem. Nas próximas seções nos concentraremos nas objeções que John Searle[41] levanta contra essa visão. A Seção 2.1 explicita o alvo das críticas de Searle. Para tal, algumas distinções importantes são apresentadas. Estaremos preocupados com os ataques que Searle faz contra o que ele chama de Cognitvismo, a tese de que o cérebro é um hardware de computador. A Seção 2.2 expõe brevemente alguns pontos importantes da teoria de computabilidade, que nos servirão de base para as discussões da Seção 2.3. Nessa última seção, apresentamos de maneira detalhada o dilema colocado por Searle. Para o autor, o Cognitivismo é incoerente em suas premissas mais fundamentais, o que compromete todo um programa de pesquisa atualmente adotado nas ciências cognitivas.

2.1 A CRÍTICA AO COGNITIVISMO

Searle (1980) critica a tese de que a mente é um programa de computador, com seu famoso argumento do quarto chinês[42]. O argumento, altamente controverso (ver Cole 2020),

[40] Turing propõe substituirmos a questão de se máquinas podem pensar pelo problema colocado pelo jogo da imitação. Deveríamos fazer a pergunta, portanto, de se máquinas serão capazes de vencer o jogo, a qual o matemático responde afirmativamente. No entanto, de acordo com a TCM, é possível que um computador vença o jogo sem empregar, de fato, inteligência genuína. Primeiro, porque os símbolos que o computador manipula só tem significado para nós, e pode ser argumentado que é uma exigência que os símbolos possuam conteúdo intrínseco. Segundo porque, embora o computador seja capaz de imitar o comportamento humano, i.e., replicar as mesmas funções em termos de input/output, ele pode vencer o jogo sem replicar os mesmos processos internos executados pelo cérebro. Isso é apenas dizer que a mera exibição de um determinado comportamento não é em si suficiente para constituir inteligência.

[41] John Rogers Searle (nascido em 1932, em Denver (EUA)) é um renomado filósofo analítico conhecido por seus trabalhos nas áreas de filosofia da linguagem, filosofia da mente e ontologia social. Leciona na Universidade de Berkeley, na Califórnia (EUA). É um ferrenho adversário das teorias que tratam o cérebro como um sistema computacional, e defende um tipo de naturalismo biológico a respeito da mente. Em 2000, foi laureado com Prêmio Jean Nicod, concedido todo ano na França a um filósofo ou cientista da mente.

[42] Searle imagina um cenário em que um indivíduo está trancado em um quarto e recebe cartas escritas em chinês, as quais ele deve responder. Apesar de não entender absolutamente nada desse idioma, ele possui um grande livro

busca despertar a intuição do leitor para a ideia de que a mera manipulação de símbolos é insuficiente para gerar intencionalidade ou entendimento (*understanding*). Esse argumento ataca o que ele mesmo chama de Inteligência Artificial Forte (IA Forte), a corrente teórica que defende que ter mente é implementar um software de computador. De acordo com a IA Forte, a simulação computacional de um processo mental (e.g. compreensão linguística) é, em si mesma, um processo mental — diferentemente de outras simulações computacionais (e.g. simulação dos movimentos planetários ou do processo de digestão do estômago), que são meras simulações — porque a simulação coincide exatamente com aquilo que define os processos mentais, viz., a execução de um programa.

Entretanto, é preciso deixar claro que a Teoria Computacional da Mente, como apresentada no Capítulo 1, não afirma que a mente é um software de computador. Ela se limita a oferecer uma hipótese a respeito dos mecanismos mentais subjacentes às transições racionais entre estados psicológicos. Há, contudo, uma objeção levantada pelo filósofo que atinge diretamente os pressupostos teóricos da TCM, qual seja, a de que o cérebro não é e não poderia ser um sistema computacional (Searle, 1992)[43]. Mas essa é exatamente uma das asserções feitas no capítulo anterior. Caracterizamos o cérebro como um sistema que instancia e transforma linhas simbólicas de acordo com as propriedades formais dos símbolos que ele manipula — precisamente o que um computador faz — e, por isso, como um sistema computacional. Essa objeção de John Searle será central nesta dissertação.

Para deixarmos claros os termos, Searle chama a tese de que a mente é o software de um computador de Inteligência Artificial Forte; a tese de que a mente pode ser *simulada* por um software de computador de Inteligência Artificial Fraca; e a de que o cérebro é um sistema computacional de Cognitivismo. O autor nega a IA Forte, mas aceita a IA Fraca. Ele afirma que, do mesmo modo que o processo digestivo que ocorre no nosso estômago ou o comportamento de um furacão podem ser simulados por um computador, a mente pode ser simulada computacionalmente. No entanto, Searle nega, contra a IA Forte, que uma simulação

---

de instruções em sua língua materna, o qual contém todas as regras necessárias para que ele forneça as respostas adequadas (respostas que um falante nativo daria). O ponto do argumento é o de que mesmo que a pessoa de fora do quarto fosse levada a acreditar que quem está dentro do quarto compreende a língua, o indivíduo continua sem entender nada de chinês. Ele simplesmente segue regras formais que o ajudam a *simular* entendimento. Mas isso é exatamente o que um computador faz. Um computador recebe informação como input (cartas em chinês), lê essa informação com um processador central (indivíduo no quarto), transforma-a utilizando regras formais (livro de instruções) e entrega um output (respostas em chinês). O computador, assim como o indivíduo no quarto, não entende o significado dos símbolos que ele manipula. As regras que ele segue dizem respeito unicamente às propriedades formais dos símbolos, não ao seu conteúdo. Searle conclui, assim, que um computador é incapaz de possuir entendimento genuíno, embora seja capaz, talvez, de simulá-lo.

[43] Quando Searle enuncia a objeção, ele afirma que o cérebro não pode ser um *computador digital*. No Capítulo 2 nós refinamos a asserção 'O cérebro é um computador' para 'O cérebro é um sistema computacional'. O que Searle entende por computador digital, porém, é o que entendemos por sistema computacional.

computacional da mente tenha qualquer propriedade psíquica. Ela pode ter valor heurístico, e.g., assim como a simulação do comportamento de um furacão nos ajuda a prever sua rota, a simulação de um processo mental pode iluminar aspectos operacionais da mente. Mas, assim como a simulação de um furacão não é um furacão e a simulação do processo digestivo não digere nada, a mera simulação de um grupamento de neurônios não gera intencionalidade, entendimento, percepção ou qualquer fenômeno de natureza mental. Devemos distinguir a coisa simulada da simulação.

Alguém poderia sugerir que o Cognitivismo e a IA Forte são teses equivalentes. Entretanto, segundo Searle, a segunda faz uma asserção mais forte do que a primeira, i.e., ela implica a primeira. Se a mente for um software, então, necessariamente, o cérebro é um sistema computacional. A inversa, por outro lado, é falsa, pelo fato de que é possível que outras qualidades – afora a qualidade de ser um sistema computacional – sejam necessárias para que um sistema seja dotado de qualidades mentais.

> Even for those who agree that programs by themselves are not constitutive of mental phenomena, there is still an important question: Granted that there is more to the mind than the syntactical operations of the digital computer, it might be the case that mental states are *at least* computational states, and mental processes are computational processes operating over the formal structure of these mental states (SEARLE 1992, p. 201).

Esses esclarecimentos são essenciais para se entender o que Searle está atacando. Nas seções seguintes, refinaremos sua objeção progressivamente, expondo suas principais linhas argumentativas. Veremos que o autor mantém uma posição completamente divorciada do paradigma computacionalista adotado em grande parte das ciências cognitivas.

2.2 A NATUREZA DA COMPUTAÇÃO

Talvez a questão mais importante colocada por Searle seja a seguinte: Que tipo de fato sobre o cérebro poderia constituir seu ser como computador? (Searle 1992, p. 204). Essa não é uma questão epistêmica, a respeito de como poderíamos *saber* que o cérebro é um computador. Antes, é uma questão metafísica, a respeito de que fatos *constituiriam* ou *determinariam* que o cérebro é um sistema computacional. Pode parecer absurdo, mas poderíamos fazer a mesma pergunta para os próprios computadores, smartphones e notebooks que usamos recorrentemente. Quais fatos sobre eles fazem com que sejam sistemas computacionais? Qual a diferença fundamental entre esses objetos e pedras, árvores e sistemas planetários (sistemas que

não computam)? Em que circunstância um sistema computa algo? Certamente, transformar inputs em outputs parece insuficiente, já que até pedras fazem isso: elas transformam os raios solares que incidem sobre ela em calor, por exemplo. Podemos, a propósito, escrever uma função matemática que descreve o calor em função dos raios e, a partir disso, criar um modelo computacional que simula o que acontece com uma pedra quando a colocamos sob a luz do Sol. Deveríamos dizer que a pedra executa esse programa? Dificilmente estaríamos dispostos a admitir que objetos como pedras e árvores computam. Para um defensor da TCM, no entanto, é preciso dizer por quê. É necessário haver algo que notebooks e cérebros fazem (mas pedras e árvores, não) que os tornam sistemas computacionais.

John Searle argumenta que a afirmação de que o cérebro computa pode ser interpretada de duas maneiras: em uma delas, a afirmação é simplesmente falsa; na segunda interpretação, ela é trivial, e o cérebro computa no mesmo sentido que pedras, árvores e sistemas planetários computam. O objetivo da Seção 2.4 é apresentar esse dilema detalhadamente. Mas para entender seus argumentos, precisamos entender um pouco a teoria de computabilidade.

2.2.1 A TESE DE CHURCH-TURING

Os primeiros computadores eram humanos. *Computador* era o nome que se dava à pessoa que realizava contas matemáticas a lápis em um papel, seguindo instruções simples e precisas. As instruções eram chamadas de *algoritmos*, que não são nada mais do que receitas que, se seguidas, fornecem o resultado desejado. Por exemplo, aprendemos no colégio primário a multiplicar algoritmicamente números com mais de um dígito. Se quisermos multiplicar 18921 por 846, há um procedimento direto que pode ser executado, o qual envolve passos simples que podem ser facilmente efetuados sem exigência alguma de criatividade ou inventividade por parte de quem os executa. Se cada passo for cumprido corretamente, o algoritmo nos fornece o resultado desejado.

Algoritmos, porém, não se restringem a operações aritméticas. Podemos criar algoritmos para trocar uma lâmpada, jogar xadrez ou dirigir um carro de um ponto inicial a um ponto final. Algoritmos são procedimentos gerais que obedecem a certos critérios parar resolver problemas. Mal'cev (1970) sugere os seguintes critérios:

(i) Um algoritmo é um procedimento de construção de quantidades através de uma medida discreta de tempo, em que cada quantidade sucessora é determinada por uma regra definida aplicada à quantidade predecessora no tempo;

(ii) a quantidade em qualquer estágio não inicial do procedimento é unicamente determinada pelo sistema de quantidades obtido nos momentos de tempo anteriores a ela;

(iii) as regras de obtenção de quantidades são simples e locais;

(iv) se o método para obter quantidades não fornecer um resultado, o algoritmo precisa dizer o que deve ser considerado como resultado;

(v) o sistema inicial de quantidades pode ser escolhido a partir de um conjunto potencialmente infinito.

Outros critérios podem ser oferecidos (ver Epstein & Carnielli 2008). Entretanto, chamamos a atenção da leitora para o fato de que essas especificações são informais e carecem de uma formulação precisa e rigorosa. Poderíamos dizer que elas tentam capturar a essência do que é um algoritmo, mas se utilizam de noções que, elas mesmas, exigem definições mais precisas.

A noção de algoritmo só foi formalizada de fato com os trabalhos de Turing e Church[44], os quais chegaram a um resultado surpreendente. Os dois oferecem definições formais do que é um processo algorítmico de maneira independente um do outro, usando noções diferentes. Mas, inesperadamente, as duas definições coincidem. Usualmente, porém, usa-se a definição de Alan Turing, embora os dois autores mereçam crédito pelos avanços teóricos.

Turing idealiza uma máquina composta de:

a. uma *fita* potencialmente infinita, dividida em quadrados que podem receber um de dois símbolos: 0 ou 1;
b. um cabeçote, que pode executar as seguintes tarefas:
    a. escanear um quadrado por vez;
    b. ler o conteúdo do quadrado escaneado;
    c. escrever 0 ou 1 no quadrado escaneado (mesmo que já haja algo escrito);
    d. movimentar-se uma vez para a esquerda (E);
    e. movimentar-se uma vez para a direita (D).

Além disso, a máquina está sempre em algum estado q_i, sendo que sua ação depende unicamente desse estado e do quadrado que está sendo escaneado. Uma instrução para essa máquina é uma quádrupla (q_i, S, A, q_j), tal que q_i é o estado atual, $S \in \{0, 1\}$, $A \in \{0, 1, E, D\}$ e

[44] 14 de junho de 1903 – 11 de agosto de 1995.

q_j é o estado subsequente. Essa quádrupla deve ser lida do seguinte modo. A máquina estará no estado q_i atual, escaneando um quadrado que terá inscrito em si 0 ou 1, e deverá executar uma das ações seguintes: escrever 0, ou escrever 1, ou ir para a esquerda ou ir para a direita. Depois, a máquina irá para o estado q_j. Um programa (ou algoritmo) é uma lista de instruções, ou seja, uma lista de quádruplas como essa. Para dar um exemplo simples, suponha que temos uma fita com uma sequência de n números 1's inscritos de maneira ininterrupta nela, e queremos apagar todos eles, i.e., escrever 0 no lugar de todo símbolo 1.

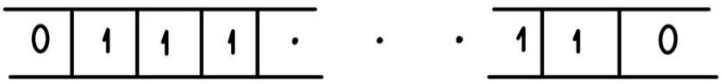

Figura 2.1 - Fita de uma máquina de Turing com uma sequência indefinida de 1's.

O que queremos é um programa que execute a função 'apagar memória'. Precisamos assumir apenas como convenção que o cabeçote da máquina começa no 1 mais à esquerda da sequência, e que o programa é encerrado se ele chegar num estado q_j que não possui instruções subsequentes. Considere o seguinte algoritmo.

(q_1, 1, 0, q_2) – *A máquina apaga o 1 escaneado e muda de estado*

(q_2, 0, D, q_1) – *Se houver mais 1's na sequência, a máquina inicia uma rotina para apagá-los da esquerda para a direita*

(q_1, 0, 0, q_3) – *Quando todos forem apagados (ou se n = 0, sendo n o número de 1's da sequência), a máquina para no estado mais alto, e o programa é encerrado.*

Existe mais de um modo de executar a função 'apagar memória'. Para ser exato, para qualquer função computável, existe um número infinito de maneiras de se executá-la. Esse é apenas um exemplo simples do que pode ser feito por uma máquina de Turing. Podemos criar instruções que dupliquem o número de 1's na fita, por exemplo. Ou, se representarmos os números em linguagem binária, podemos realizar operações aritméticas sobre eles. Somá-los, subtraí-los, dizer se um número é par ou ímpar, se é primo ou não etc. Na verdade, qualquer operação que um computador moderno executa pode ser executada por essa máquina simples. Basta criar um modo de representar em linguagem binária as informações que precisam ser processadas e elaborar uma lista de instruções que levem a máquina a manipular os símbolos de

modo a executar a função desejada, como fizemos no exemplo acima. Alguns teóricos se arriscam a dizer que qualquer função executada por qualquer computador possível pode ser executada por uma Máquina de Turing. Aceitar essa tese é aceitar a *Tese de Church-Turing*. A Tese de Church-Turing diz que uma função é efetivamente computável se, e somente se, ela for computável por uma máquina de Turing (Biraben (1994), (2001)).

A Máquina de Turing é um dispositivo teórico formal que tenta capturar a noção informal de algoritmo. A Tese de Church-Turing não pode ser provada, justamente por tentar estabelecer uma relação de equivalência entre uma noção formal e uma noção informal. Aqui, porém, entra o resultado surpreendente que corrobora com ela. Houve outras tentativas de formalizar a noção do que é efetivamente computável. Todos os modelos, no entanto, se mostraram exatamente equivalentes. Turing (1936) demonstra a equivalência entre a Máquina de Turing e as sequências λ-definíveis de Alonzo Church. Ambas, por sua vez, são equivalentes à máquina abstrata de Emil Post[45], às funções recursivas de Kurt Gödel e aos algoritmos de Markov[46] (ver Epstein & Carnielli 2008; Immerman 2015). O fato de que todos esses modelos teóricos diferentes delimitam uma mesma classe de objetos parece sugerir, ao menos, que a noção informal de algoritmo corresponde, de fato, a uma categoria matemática bem definida.

A vantagem do modelo criado por Turing é a de que ele torna relativamente fácil de entendermos quando um sistema concreto está realizando uma computação. Seu modelo matemático é, simultaneamente, um projeto de engenharia de uma máquina autônoma que implementa tarefas algoritmicamente. Basta seguirmos suas instruções para criarmos tal máquina. Precisamos de um aparelho que imprima e manipule 0's e 1's numa fita de acordo com regras formais, i.e., regras que respeitem as formas dos símbolos sem se preocupar com o conteúdo deles, embora respeitando suas relações semânticas. Entretanto, é importante notar que nem a fita e nem os símbolos são essenciais para o funcionamento da máquina, apenas *objetos que cumpram o mesmo papel que eles*. Computadores modernos são espécies de máquinas de Turing. Porém, se os abrirmos, não veremos 0's e 1's inscritos numa fita quadriculada. Os símbolos são codificados de outras maneiras, como em polaridades magnéticas numa placa de silício.

Não é necessário também que os símbolos primitivos pertençam ao conjunto {0, 1}; uma máquina poderia manipular símbolos de 0 a 10, ou letras do alfabeto, ou palavras inteiras, por exemplo. O sistema binário, porém, é a linguagem mais econômica possível em termos simbólicos, o que permite um design mais eficiente do hardware do computador. Suponha que,

[45] 11 de fevereiro de 1897 – 21 de abril de 1954.
[46] 14 de junho de 1856 – 20 de julho de 1922.

ao invés de manipular 0's e 1's, nossos computadores manipulassem 0's, 1's, 2's, ..., 9's e 10's. Teríamos que arquitetar onze tipos de estados físicos diferentes para codificar cada um desses símbolos. Não seria possível implementar essa linguagem apenas com polaridades positivas e negativas, ligado e desligado. A máquina teria que ser capaz de distinguir entre onze estados físicos distintos, talvez com onze níveis de voltagem diferentes, por exemplo. Essa máquina teria uma arquitetura interna mais complexa e, portanto, um custo de produção maior (é importante ter em mente que o modo como computadores atuais são fabricados está sujeito às demandas do mercado, e não porque ele é melhor de um ponto de vista teórico, por exemplo). Além disso, essa arquitetura seria mais suscetível a erros de leitura, já que o cabeçote teria que discriminar não entre estados "opostos" (ligado/desligado), mas entre uma sucessão de estados próximos entre si.

Ademais, a Tese de Church-Turing nos diz que o sistema simbólico que um computador utiliza é indiferente do ponto de vista de seu poder computacional. Uma máquina que utilize algarismos de 0 a 10 não é mais poderosa do que outra que utilize linguagem binária. Pode haver diferenças no custo computacional de se realizar uma determinada tarefa, mas ambas serão capazes de realizá-la. A respeito disso, Shanon (1993) prova que qualquer máquina de Turing que implementa um sistema simbólico com n símbolos pode ser simulada por uma máquina de Turing com apenas dois símbolos.

A este ponto, uma questão curiosa se apresenta: computação envolve, necessariamente, manipulação simbólica? Alguns autores (ver Copeland 1993, p. 310) definem o termo *computador* como um manipulador simbólico. Outros autores (ver Rescorla 2019; Thagard 2018) aplicam o termo *computar* mesmo para sistemas processadores de informação que não manipulam símbolos, como redes conexionistas. Essa questão parece ser terminológica. Não há nenhuma restrição explícita na teoria de computabilidade que estabeleça uma resposta definitiva. De acordo com Rescorla (2020), sistemas computacionais clássicos (à la máquinas de Turing) podem implementar sistemas conexionistas, e sistemas conexionistas podem implementar sistemas computacionais clássicos. Não vemos nenhuma razão para sermos restritivos quanto ao termo *computação*. Entretanto, para esta dissertação, o sistema computacional relevante é aquele que opera sobre estruturas simbólicas.

2.3 DUAS INTERPRETAÇÕES DO COGNITIVISMO

Podemos voltar à pergunta colocada por John Searle: que tipos de fatos sobre o cérebro poderiam determinar ou constituir seu ser como um sistema computacional? Antes de responder

a essa pergunta, é construtivo pensar no que faz com que aparelhos eletrônicos como notebooks e smartphones sejam sistemas computacionais. Como foi dito acima, se abrirmos esses aparelhos não encontraremos 0's e 1's inscritos numa fita. Computadores atuais codificam 0's e 1's por meio de polaridades magnéticas inscritas em placas de silício. Não há nada de especial, porém, nesses substratos físicos. As razões pelas quais nossos computadores são construídos dessa forma envolvem fatores como eficácia computacional e viabilidade econômica de produção. Mas um computador feito de engrenagens e alavancas seria capaz, em princípio, de executar as mesmas funções. Assim, o material do qual é feito um sistema computacional é irrelevante no que diz respeito ao que faz com que ele seja um sistema computacional. Poderia ser feito de silício, de madeira, ou mesmo de carne e osso.

Uma vez que sistemas computacionais são multiplamente realizáveis, é natural qualificá-los como objetos funcionais, assim como carburadores e termostatos. Mas há uma diferença crucial entre eles. Dizemos que carburadores e termostatos são multiplamente realizáveis porque podemos construí-los a partir de uma variedade de substratos físicos diferentes que conseguem executar a mesma tarefa. É necessário que o substrato físico utilizado possua as propriedades causais que caracterizam as funcionalidades desses artefatos. Lembre-se de que a doutrina funcionalista individua objetos funcionais de acordo com a função que eles cumprem num determinado sistema. Isso significa que a realizabilidade múltipla de carburadores e termostatos ainda é delimitada de alguma forma. Não podemos construir um carburador feito de isopor ou um termostato feito de queijo porque esses materiais são inadequados para implementar as funções desejadas. A realizabilidade múltipla de estados computacionais, por outro lado, é ilimitada, pois ela não deriva do fato de que 0's e 1's possuem tais e tais poderes causais, mas do fato de que *0's e 1's podem ser atribuídos a diferentes substratos físicos ao nosso bel prazer*.

> The classes of carburetors and thermostats are defined in terms of the production of certain *physical* effects. That is why, for example, nobody says you can make carburetors out of pigeons. But the class of computers is defined syntactically in terms of the *assignment* of 0's and 1's. The multiple realizability is a consequence not of the fact that the same physical effect can be achieved by different physical substances, but the relevant properties are purely syntactical. The physics is irrelevant except in so far as it admits of the assignments of 0's and 1's and of state transitions between them (SEARLE 1992, p. 209).

Notebooks e smartphones são sistemas computacionais porque eles foram projetados para articular estados internos que representam símbolos. Mas eles manipulam símbolos porque nós os *interpretamos* dessa forma. Sistemas computacionais não poderiam existir na natureza do mesmo modo que árvores e rios existem na natureza, porque símbolos possuem uma natureza

sintática, e propriedades sintáticas não correspondem a propriedades físicas, mas a abstrações que seres inteligentes estabelecem quando criam sistemas linguísticos. O termo *símbolo* não nomeia objetos físicos, mas convenções humanas a respeito de objetos físicos. Disso se segue que a propriedade de computar algo concretamente é, também, uma espécie de convenção humana e, portanto, relativa a um observador ou usuário que interpreta os estados internos de um sistema de um modo específico, de acordo com seus próprios interesses.

Alguns autores afirmam que a tese de que o cérebro computa é um problema empírico (Fodor 1975; Pylyshyn 1984, Copeland 1996). Mas devemos nos perguntar: seria essa tese suscetível de confirmação ou desconfirmação empírica, do mesmo modo que podemos confirmar ou desconfirmar empiricamente a tese de que plantas fazem fotossíntese, ou de que humanos e chimpanzés possuem um ancestral comum? Que tipo de fato sobre o cérebro instituiria que ele é um sistema computacional? Se Searle estiver correto, não há nenhum fato objetivo que estabeleça, independentemente da interpretação e perspectiva de terceiros, se o cérebro está computando ou não. É impossível *descobrir* que o cérebro é um computador, porque ser um computador não é uma propriedade intrínseca a nenhum objeto[47].

Se Searle estiver correto, portanto, todas as explicações fundamentadas em procedimentos computacionais a respeito das operações cognitivas que o cérebro realiza são falsas. Elas são falsas porque as explicações oferecidas nas ciências cognitivas têm a pretensão de serem *causais*. Por exemplo, gostaríamos de dizer que é em virtude dos procedimentos computacionais A, B e C que expressamos a habilidade cognitiva H. Entretanto, de acordo com o autor, o cérebro não implementa A, B e C, se não nos olhos de um observador. Mas, é claro, a habilidade cognitiva H não ocorre apenas quando há um observador atribuindo interpretações computacionais aos processos neurais dentro de nossos crânios. A habilidade cognitiva H, seja ela qual for, existe desde muito antes de qualquer teoria da computabilidade, antes da existência de qualquer símbolo.

O que Searle está afirmando não é de modo algum absurdo. O próprio pai da computabilidade, Alan Turing, afirma com todas as letras que não existem sistemas computacionais concretos, apenas *objetos que poderiam ser vistos dessa forma*.

---

[47] *Ser um símbolo* é uma propriedade relativa a um observador do mesmo modo que *ser uma cadeira* é uma propriedade relativa a um observador (Searle 1992, p. 211). Não é possível *descobrir* cadeiras na natureza. É possível descobrir objetos com a forma de uma cadeira, e que seriam adequados para esse fim. Mas algo só é uma cadeira a partir do momento em que alguém o utiliza como tal. Do mesmo modo, algo só é um símbolo relativamente a alguém que o utiliza desse modo. Se todas as pessoas desaparecessem, o mundo não conteria mais cadeiras, símbolos, moedas e contratos empresariais; mas conteria árvores, nuvens, rios e minerais. Searle distingue, assim, entre propriedades que são relativas a um observador/usuário e propriedades que são intrínsecas ao objeto. O autor argumenta que nenhuma ciência natural fornece suas explicações em termos de propriedades relativas a um observador (Searle 1992, p. 212).

> The digital computers considered in the last section may be classified amongst the 'discrete state machines'. These are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously. But there are many kinds of machine which can profitably be *thought of* as being discrete state machines. For instance in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for must purposes we can forget about them (TURING 1950, p. 439).

Embora Searle e Turing neguem a existência de máquinas computacionais concretas por motivos diferentes – enquanto um afirma que *símbolos* não existem, o outro afirma que *processos discretos* não existem –, acredito que a ideia implícita em ambos os posicionamentos seja a mesma. Máquinas de Turing não existem no mundo físico pelo mesmo motivo que triângulos e esferas não existem no mundo físico, apenas coisas que podem ser tratadas como triângulos e esferas. Máquinas de Turing, assim como objetos geométricos, pertencem ao mundo abstrato da matemática e, independentemente da posição ontológica que a leitora mantenha com relação a esses objetos, é ponto pacífico que eles não existem do mesmo modo que rios e árvores existem. Mas, assim como é frequentemente conveniente tratar objetos físicos como se fossem triângulos e esferas perfeitas (pela misteriosa razão de que objetos físicos frequentemente se comportam como objetos matemáticos), também é útil tratar objetos como se fossem máquinas de Turing. É preciso ter em mente, porém, que são apenas maneiras pragmáticas de tratar objetos físicos, pois *computação*, *algoritmo* e *símbolo* são noções matemáticas e, portanto, abstratas por excelência.

Alguém poderia retorquir que essa é uma maneira muito "materialista" de ver as coisas; que se todos os seres humanos desaparecessem e, por acaso, um macaco encontrasse minha calculadora e apertasse os botões "5", "+" e "7", ela computaria efetivamente "12", mesmo sem haver alguém naquele momento capaz de interpretar seus estados internos. Entretanto, Searle argumenta, isso é um engano. Seu comportamento de alto nível pode ser uma simulação engenhosa de raciocínios matemáticos (para nós que ocupamos o papel de observadores nesse experimento mental), mas, no fundo, nenhuma calculadora sabe realizar contas de verdade. Pensamos que ela somaria 5 e 7 para calcular 12, mas o que ela faria primeiro seria transcrever esses números para a notação binária – ela não sabe somar em bases de 10. Depois, ela precisaria aplicar um algoritmo que some números binários. Esse algoritmo, porém, não diz nada a respeito de como *somar* números. Eles serão algo mais similar às quádruplas que instruem uma máquina de Turing a mover e apagar 0's e 1's. No fundo, então, a calculadora não entende o conceito de

somar quantidades, apenas o de mover 0's e 1's de lugar. Bem, na verdade, a calculadora também não sabe o que são 0's e 1's, pois não há nada dentro dela que se pareça com esses símbolos. O que ela faz é modular certos sinais elétricos que um dia alguém engenhosamente utilizou para codificar 0's e 1's. Através desse método de decomposição recursiva (ver Searle 1992, p. 213), é possível ver que dentro da calculadora não há nenhum processo inteligente, apenas operações mecânicas realizadas cegamente pelos componentes eletrônicos que compõem o aparelho.

Essa é a primeira interpretação de Searle a respeito da tese de que o cérebro computa. O cérebro computa apenas do ponto de vista de um observador, porque os elementos básicos da computação, viz. símbolos, são atribuídos a objetos físicos de acordo com o olhar e interesse de alguém que os interprete dessa maneira. Evidentemente, não é nesse sentido que cientistas cognitivos afirmam que o cérebro computa. Precisamos de uma noção mais robusta de computação para sustentar o programa cognitivista.

Searle considera uma corrente teórica alternativa, segundo a qual sistemas computacionais concretos existem e operam de maneira independente de qualquer observador ou usuário que confira interpretações simbólicas aos seus estados internos. De acordo com essa corrente, outros elementos que não o olhar interessado de um ser inteligente fixam quais operações computacionais um sistema implementa. A atribuição de um processo algorítmico a um sistema concreto deve ser entendida como condição *suficiente* para dizer que ele computa, mas não *necessária*. Uma máquina M computa um programa C se, e somente se, houver uma correspondência entre os estados internos de M e os estados especificados por C. Se essa correspondência existir, o programa C descreve o comportamento de M e M implementa C concretamente. Não é necessário que um observador atribua a interpretação C à máquina M. Não é necessário sequer que alguém esteja ciente de que C é uma descrição computacional de M.

Mais precisamente, essa correspondência deve atender a dois pontos: a) há um mapeamento entre, de um lado, os estados internos de M descritos fisicamente e, de outro, os estados computacionais especificados em C, tal que b) as transições entre os estados internos de M espelhem as transições entre estados computacionais especificados em C (ver Piccinini (2017)). Essa visão foi defendida por Putnam (1975a) e é conhecida como *Simple Mapping Account* (SMA, Piccinini (2017)).

De acordo com essa visão, não é a atribuição de estados computacionais a partes internas de um objeto que fazem com que ele implemente um programa, mas a existência de um isomorfismo entre seus estados internos e os estados abstratos de um algoritmo. Em princípio,

portanto, se for possível interpretar objetos computacionalmente, é porque esses objetos estão, de fato, executando um programa. A conclusão inescapável aqui é a de que tudo computa. Se eu interpretar potes de manteiga como 1 e potes de requeijão como 0, minha mesa de café da manhã armazena o número 2. Com efeito, minha mesa de café da manhã implementa muitos estados computacionais simultaneamente, pois existem várias maneiras de interpretar os itens que estão em cima dela. O que dizer do cérebro, que possui cerca de 86 bilhões de neurônios? E da nossa galáxia, que possui mais de 200 bilhões de estrelas?

A hipótese cognitivista se coloca distante de qualquer confirmação ou desconfirmação empírica, já que a noção de computação concreta se trivializa.

> Well, we wanted to know how the brain works, specifically how it produces mental phenomena. And it would not answer that question to be told that the brain is a digital computer in the sense that stomach, liver, heart, solar system, and the state of Kansas are all digital computers. The model we had was that we might discover some fact about the operation of the brain that would show that it is a computer. We wanted to know if there was not some sense in which brains were *intrinsically* digital computers in a way that green leaves intrinsically perform photosynthesis or hearts intrinsically pump blood. And what we were asking is, "Is there in that way a fact of the matter about brains that would make them digital computers?" It does not answer that question to be told, yes, brains are digital computers because everything is a digital computer (SEARLE 1992, p. 208).

Assumindo a SMA como modelo de computação concreta, o cognitivismo corre o risco, além disso, de trivializar processos mentais.

> On the standard textbook definition of computation, it is hard to see how to avoid the following results:
> 1. For any object there is some description of that object such that under that description the object is a digital computer.
> 2. For any program and for any sufficiently complex object, there is some description of the object under which it is implementing the program. Thus for example the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar, then if it is a big enough wall it is implementing any program, including any program implemented in the brain (Idem).

O ponto 2 acima colocado por Searle é o que Jack Copeland chama de *Teorema de Searle* (Copeland 1996, p. 344). Em seu artigo de 1996, *What is computation*, ele oferece uma prova desse teorema, demonstrando que para qualquer programa de computador α e para qualquer objeto O com um número suficientemente grande de partes internas, é possível estabelecer uma relação de correspondência entre os estados computacionais de α e os estados

internos de O, tal que as transições entre os estados internos de O espelhem as transições dos estados computacionais de α. A seguir, apresentamos a prova de maneira esquemática.

É preciso definir alguns termos. Um algoritmo, como vimos, consiste numa lista de instruções (chamemos α) que, se seguidas, executam uma função f. Isto é, α toma como input argumentos de f e devolve como output valores de f. Diremos que α é específico a uma arquitetura particular, o que significa que esse algoritmo recorre a certos símbolos primitivos e aplica certas operações específicas. Por exemplo, a arquitetura das máquinas de Turing utiliza 0's e 1's como símbolos primitivos e as operações consistem em apagar e escrever 0's e 1's e levar o cabeçote para a direita ou para a esquerda. Um algoritmo destinado a máquinas de Turing não pode ordenar que a máquina *some* 1's porque ela não reconhece essa ação. Por outro lado, podemos ter arquiteturas diferentes, que utilizam mais símbolos e também podem aplicar outras operações, como soma ou multiplicação[48]. Copeland chama de SPEC o algoritmo α junto com uma especificação de sua arquitetura.

A ponte entre uma lista abstrata de instruções e um objeto concreto O ocorre através de um *esquema de marcação L*. O esquema de marcação estabelece (1) certas partes da máquina como portadoras simbólicas e (2) um método para descobrir qual símbolo uma portadora simbólica expressa em qualquer momento no tempo. *L* constitui um código para que seja possível interpretar semanticamente os estados físicos da máquina. Por exemplo, um esquema de marcação possível pode estabelecer que estados voltaicos entre 0V e 1V representam 0 e estados voltaicos maiores que 1V representam 1.

Se houver uma relação de isomorfismo entre um objeto O sob um esquema de marcação *L* e um conjunto de instruções específicos a uma arquitetura SPEC, diremos que o par ⟨O, *L*⟩ é um modelo de SPEC. A Simple Mapping Account pode, então, ser formulada do seguinte modo:

Simple Mapping Account: Um objeto O computa uma função f se, e somente se, existe um esquema de marcação *L* e uma especificação formal SPEC (de uma arquitetura e um algoritmo que tome como input argumentos de f e entregue valores de f como output) tal que ⟨O, *L*⟩ é modelo de SPEC[49].

[48] Como vimos na seção 2.3, isso não tornaria a máquina mais poderosa. Estamos apenas tornando explícito novamente a possibilidade de haver outras arquiteturas computacionais.

[49] A asserção "existe um esquema de marcação L..." deve ser entendida do mesmo modo que, em matemática, entendemos "existe uma função....". Não é preciso que alguém conheça o esquema de marcação (ou a função) para que a afirmação existencial seja verdadeira.

O Teorema de Searle, assim, é expresso como se segue:

Teorema de Searle: Para qualquer entidade x (com um número suficientemente grande de partes discrimináveis) e para qualquer algoritmo y específico a uma arquitetura, há um esquema de marcação *L* tal que ⟨x, *L*⟩ é modelo de y.

Para provar o teorema, Copeland considera um programa aleatório – como uma sessão do Wordstar – e um objeto grande o suficiente para que haja um número confortável de estados internos discrimináveis que sirvam de portadoras simbólicas – como a parede que Searle menciona em seu argumento. Como nenhuma parte da prova faz uso essencial do Wordstar ou da parede de Searle, o resultado se generaliza para qualquer objeto O e qualquer programa α.

O esquema de marcação é feito do seguinte modo. Assumindo que a parede é constituída de polímeros, uma região R da parede instancia 1 se o número de polímeros contidos nela for par, e instancia 0 se o número de polímeros for ímpar.

Os registros, digamos, consistem em fórmulas de 8 bits. Uma instrução I possui a forma 00000001, por exemplo. O programa deve especificar o que cada instrução faz. Exemplo:

I = 00000001 ACTION-IS (M = 00000000)

O comportamento do programa é descrito pelo termo ACTION-IS, que significa simplesmente que a instrução antecedente determina que o programa execute a função consequente (no caso, M = 00000000). Suponha que M = 00000000 seja a operação "apagar memória" ou "o conteúdo da memória se torna 00000000". Assim, a instrução acima deve ser lida como o condicional "se a instrução 00000001 for executada, então o programa apaga o que está na memória".

Suponha que, numa determinada sessão do Wordstar, uma região r da parede instancie o conteúdo M da memória. Isso não garante que r será sempre a região que corresponde a M, por duas razões. Em primeiro lugar, queremos que o conteúdo da memória permaneça constante, ao não ser que seja alterado por alguma instrução do programa, e não há garantia alguma de que as propriedades físicas de r permanecerão constantes pelo tempo requerido. Em contrapartida, queremos que o portador simbólico mude de configuração conforme as instruções do programa, e as chances de que as propriedades físicas relevantes da região r da parede acompanhem essas mudanças é praticamente nula. Portanto, a memória M não terá um referente

fixo (nem as instruções e nem qualquer outro componente operacional). Como resolvemos esse problema?

Um programa opera em ciclos. Uma sessão qualquer do Wordstar possuirá n ciclos (C_1, ..., C_n), sendo que em cada um deles a configuração do programa muda, i.e., os conteúdos das instruções, da memória etc., mudam. Para cada C_i, portanto, haverá uma instrução I_i, uma memória M_i etc. correspondentes, caracterizadas em sequências binárias de 8 bits (e.g. em C_{57} podemos ter I_{57} = 01001011 e M_{57} = 11110000). Para completar o esquema de marcação, teremos que assumir uma função cujo domínio são os ciclos do programa e cuja imagem são as regiões da parede que satisfazem as propriedades físicas adequadas para serem portadoras simbólicas. Assim, enquanto que M_{57} denota uma sequência abstrata de bits num determinado ciclo do programa, $|M_{57}|$ denota a região da parede que a codifica nesse instante (o que Copeland chamaria de 'codificação molecular do número binário M_{57}'). Diremos que $\mathbf{M}$ é a função cujo domínio é o conjunto C dos ciclos do programa e cuja imagem é o conjunto $|M|$, e o mesmo vale para $\mathbf{I}$. Copeland chama esse procedimento de *Searlificação*.

A instrução I = 00000001 ACTION-IS (M = 00000000) será interpretada como (o quantificador varia sobre instantes do ciclo):

$$\forall j\ (\mathbf{I}(j) = |00000001| \rightarrow \mathbf{M}(j) = |00000000|)$$

O que a sentença acima diz é que sempre que o referente da instrução (sob o esquema de marcação L definido acima) represente o código binário 00000001, o referente da memória (sob o esquema de marcação L definido acima) representa 00000000. Assumindo, portanto, que a parede é grande o suficiente, haverá um esquema de marcação L tal que a interpretação acima será verdadeira sobre ela. Uma vez que nenhuma parte da prova faz uso essencial das particularidades do programa e nem do objeto, o resultado pode ser generalizado.

As razões pelas quais o Teorema de Searle é preocupante para a Teoria Computacional da Mente devem ser óbvias. Uma hipótese que tenta descrever uma habilidade cognitiva qualquer em termos de um algoritmo que o cérebro implementa perde força explanatória, já que uma parede executa as mesmas computações e (plausivelmente) não possui a mesma capacidade mental que nós. Searle coloca um dilema para os defensores do Cognitivismo. Existem duas maneiras de entender computações concretas, e ambas as narrativas nos levam a becos sem saída. Ou a propriedade de computar algo é relativa a um observador e, portanto, inútil para explicar nossos processos mentais, já que eles existem desde antes de qualquer observador poder atribuir

interpretações computacionais ao cérebro; ou a propriedade de computar algo é trivial, onipresente na natureza, o que nos levaria a crer que paredes, rios e montanhas instanciam as mesmas atividades cognitivas que o cérebro (assumindo que a cognição é redutível a padrões de organização que podem ser descritos algoritmicamente). A redução ao absurdo, argumenta Searle, mostra que há algo errado com as suposições basilares das ciências cognitivas. A premissa que deve ser abandonada é a de que o cérebro é um hardware que opera sobre cadeias simbólicas.

No próximo capítulo, avaliaremos de maneira crítica os desafios lançados por Searle, na tentativa de tornar consistente a tese de que o cérebro é intrinsicamente um sistema computacional.

3. O FALSO DILEMA DE JOHN SEARLE

Vimos no capítulo anterior que Searle (1992) coloca um dilema para a Teoria Computacional da Mente. De um lado do dilema, se processos computacionais concretos forem implementados pelo cérebro em virtude de haver uma relação de isomorfismo estrutural entre nossos estados neurofisiológicos e estados algorítmicos abstratamente especificados, será uma consequência inaceitável disso, aponta Searle, que qualquer objeto computa algum programa, visto que qualquer estado físico é estruturalmente isomórfico a algum algoritmo. Em especial, uma parede grande o suficiente computaria todos os programas implementados pelo meu cérebro e, portanto, instanciaria estados cognitivos de mesmo tipo que os meus. Do outro lado do dilema, se processos computacionais concretos existirem apenas nos olhos de um observador, o qual atribuiria interpretações sintáticas às propriedades físicas de um objeto, que seria tomado como sistema computacional, então esse objeto não computaria nada de fato, pois as propriedades sintáticas que num momento lhes são concedidas não lhes são intrínsecas, mas inteiramente dependentes do olhar interessado de um intelecto externo; o máximo que poderíamos afirmar é que esse objeto se comporta *como se computasse* determinada função; a consequência inaceitável disso (para as ciências cognitivas), aponta o autor, é que explicações baseadas em termos de sintaxe, estados ou processos computacionais são falsas, espúrias ou inadequadas, pois cérebros executam atividades cognitivas independentemente da atribuição computacional de terceiros.

Neste capítulo, pretendemos argumentar que o problema colocado por John Searle constitui um falso dilema, e que há uma interpretação da Teoria Computacional da Mente e, mais precisamente, da dimensão formal dos processos cerebrais assumidos pela teoria, bastante natural e coerente com os tipos de abordagens explicativas que vemos em outras ciências especiais.

A Seção 3.1 avança a abordagem semântica de computação, que coloca restrições sobre os tipos de objetos que podem constituir sistemas computacionais. De acordo com essa visão, apenas objetos com propriedades semânticas podem ser candidatos a sistemas computacionais, evitando-se, assim, o pancomputacionalismo – uma das vias do dilema de Searle.

A Seção 3.2 introduz a ideia de conteúdo semântico e seu papel no computacionalismo clássico.

A Seção 3.3 ataca a outra via do dilema se Searle, de acordo com a qual a realizabilidade múltipla de estados computacionais advém do fato de que eles são aleatoriamente atribuídos de

acordo com o interesse de um observador. Argumentamos que essa atribuição não pode ser aleatória se o sistema possuir propriedades semânticas intrínsecas.

Por fim, a Seção 3.4 apresenta uma visão alternativa à abordagem semântica aqui defendida.

3.1 A NATUREZA SEMÂNTICA DOS PROCESSOS COMPUTACIONAIS

Ao contrário dos dois capítulos anteriores, de caráter mais expositivo que argumentativo, este capítulo expõe uma série de objeções ao modo como John Searle interpreta os fundamentos teóricos da TCM. A primeira grande divergência que temos com relação ao pensamento de Searle reside na noção que ele apresenta de computação. De acordo com o autor (Searle 1992, p. 205), a própria definição de computação oferecida por Turing (1936) descreve processos computacionais como operações formais sobre estruturas sintáticas[50]. É irrelevante, de acordo com essa definição, se essas estruturas sintáticas possuem conteúdo ou não. Seguindo Fodor (1981)[51], porém, defendemos uma visão semântica de computação, de acordo com a qual apenas entidades com conteúdo semântico são candidatas a compor estados/processos computacionais. Não argumentaremos que a definição de Turing deva ser interpretada a partir de um viés semântico. Acreditamos que a definição fornecida pelo matemático é silenciosa a esse respeito. Provavelmente, Turing estava mais preocupado com o problema de definir máquinas que *funcionem* como sistemas computacionais (que podem ser *interpretadas* dessa forma) do que com o problema de quais condições seriam necessárias e suficientes para que um sistema concreto implementasse funções algorítmicas. Mas acreditamos que sua definição é compatível tanto com a visão semântica quanto com a visão formal de computação esboçada por Searle.

Diferentemente da visão sintática de computação, a visão semântica exige que um algoritmo seja implementado sobre *estruturas representacionais*. Computações, portanto, não ocorrem sobre símbolos sem sentido, como risquinhos ou bolinhas, mas sobre símbolos com algum conteúdo, como 0's e 1's. Uma das motivações da visão semântica reside na intuição de que sempre deve haver uma resposta para a pergunta "O que esse sistema está computando?", e não é uma resposta satisfatória para essa pergunta dizer que o sistema adiciona três risquinhos a mais no fim da sequência de outros quatro risquinhos. É claro, pode haver um *nível de descrição*

[50] "[...] the class of computers is defined syntactically in terms of the *assignment* of 0's and 1's" (Searle 1992, p. 207). "If computation is defined in terms of the assignment of syntax, then everything would be a digital computer, because any object whatever could have syntactical ascriptions made to it" (Idem).

[51] "There is no computation without representation" (Fodor 1981, p. 7). Outros autores que também defendem uma visão semântica de computação são Rescorla (2017a, 2017b) e Shagrir (2018).

do sistema no qual ele simplesmente adiciona risquinhos a uma fita. Podemos nos abstrair do significado dos símbolos manipulados para analisar, compreender ou evidenciar aspectos formais dos processos implementados. Mas, novamente, é uma condição necessária da visão semântica de computação que os itens manipulados por um sistema computacional carreguem conteúdo semântico.

A visão sintática de computação não nos leva necessariamente a um pancomputacionalismo (tese de que tudo computa). Isso dependerá, dentre outras coisas, de como entendemos o termo *sintaxe*. A Simple Mapping Account (ver Capítulo 2) aceita uma interpretação extremamente liberal desse termo, de modo que é difícil imaginar uma propriedade que não possa ser interpretada como sintática. Desse modo, há uma explosão no número de sistemas computacionais, já que qualquer arranjo físico e quaisquer transições entre estados físicos são estruturalmente isomórficas a algum algoritmo. Trivialmente, uma parede grande o suficiente instanciará um padrão molecular isomórfico a qualquer descrição computacional implementada pelo meu cérebro, mas isso está longe de implicar que a parede literalmente execute algum programa.

A condição semântica restringe drasticamente o conjunto de sistemas computacionais concretos possíveis. De acordo com a visão semântica, simplesmente não se segue que a parede executa algum programa só porque é possível encontrar padrões moleculares isomórficos a algoritmos, porque seus estados internos não possuem conteúdo algum. Essa restrição não constitui nenhum absurdo, e Searle está preparado para admitir que a Simple Mapping Account pode ser liberal demais, talvez uma consequência inesperada das definições de Turing.

> I do not think the problem of universal realizability is a serious one. I think it is possible to block the result of universal realizability by tightening up our definition of computation. Certainly we ought to respect the fact that programmers and engineers regard it as a quirk of Turing's original definitions and not as a real feature of computation (SEARLE 1992, p. 209).

Além disso, em alguns momentos Searle flerta com a posição semântica aqui defendida, sugerindo a importância de haver conteúdo intencional intrínseco ao sistema para que seja possível a implementação legítima de algum processo computacional.

> In Turing's human computer there really is a program level intrinsic to the system, and it is functioning causally at that level to convert input to output. This is because the human is consciously following the rules for doing a certain computation, and this causally explain his performance. But when we program the mechanical computer to perform the same computation, the assignment of a computational interpretation is now relative to us, the outside homunculi.

> There is no intentional causation intrinsic to the system [...] *It could not be following rules because there is no intentional content intrinsic to the system that is functioning causally to produce the behavior* (SEARLE 1992, p. 216, grifo nosso)

Essa passagem é interessante também por outra razão. Searle afirma explicitamente que nós, humanos, implementamos algoritmos de maneira concreta e literalmente (ao menos quando fazemos isso conscientemente). Ele declara que existe um "nível de programa" intrínseco ao sistema, porque as regras são conscientemente representadas, e o conteúdo intencional das regras funciona causalmente para produzir um comportamento específico. Não fica bem claro o que Searle quer dizer com "existe um nível de programa intrínseco ao sistema", pois, como vimos, o autor defende uma visão sintática de computação e, além disso, argumenta que sintaxe não é uma qualidade intrínseca a nenhum objeto físico, o que parece implicar na impossibilidade de existir um sistema computacional concreto.

No momento, porém, queremos apenas ressaltar que um dos caminhos do dilema levantado por Searle — o pancomputacionalismo — não surge como um problema para a vertente da TCM que estamos defendendo. Como a visão semântica de computação requer que os símbolos manipulados numa função algorítmica concreta carreguem conteúdo semântico, poucos objetos são realmente qualificados para cumprirem esse papel.

3.2 PROPRIEDADES SEMÂNTICAS

Embora estejamos defendendo uma visão semântica de computação, reconhecemos o caráter acirrado das disputas que envolvem essa noção. De um ponto de vista filosófico, não há qualquer consenso sobre o que sejam propriedades semânticas, se elas existem e como podemos caracterizá-las. Autores diferentes sugeriram formas distintas de entender a natureza da semântica como significado, e a intuição cumpre um grande papel nesse processo. De fato, a principal razão para aceitarmos a existência de objetos com propriedades semânticas no catálogo ontológico da Teoria Computacional da Mente consiste no fato de que nossas melhores ciências a respeito do funcionamento da mente aceitam esses objetos (Fodor (1975), Burge (2010)). Não obstante, situaremos brevemente a noção de semântica adotada pela TCM frente a algumas das grandes questões filosóficas que norteiam as teorias semânticas das últimas décadas.

Uma das ideias essencialmente subjacentes ao computacionalismo clássico é o atomismo semântico. De acordo com essa tese, devemos entender nossos estados representacionais como sendo estruturados a partir de primitivos atômicos, as unidades básicas de nosso sistema

representacional. Suponha que *vermelho* e *círculo* sejam primitivos representacionais de nosso sistema linguístico interno. Significa que eles não são decomponíveis em unidades de significado mais básicas. Por outro lado, a ideia composta *círculo vermelho* é uma estrutura molecular, decomponível em unidades menores de significado. Essa tese será desenvolvida em algum detalhe mais adiante.

Outra questão importante diz respeito ao modo como individuamos estados mentais de indivíduos. Nesse ponto, divergimos de Fodor (1987) quanto à importância do contexto e ambiente no qual o organismo está inserido. Fodor defende que estruturas físicas qualitativamente idênticas não podem implementar computações diferentes. A ideia é que quais computações um sistema está executando dependerá unicamente de suas propriedades locais, intrínsecas a ele. Argumentaremos a favor da ideia de que computações são individuadas (ao menos parcialmente) em função da informação que está sendo processada e, ao mesmo tempo, que o conteúdo informacional processado por um sistema é individuado com relação ao ambiente no qual ele se encontra e com o qual possui contato direto.

3.2.1 ATOMISMO SEMÂNTICO

Vimos no Capítulo 2 (§ 2.5.2) que, segundo a hipótese da linguagem do pensamento, há um isomorfismo entre o conteúdo intencional de uma atitude proposicional e o veículo físico que corresponde a essa atitude. Isso significa que se João possui a crença 'Alienígenas gostam de bolo de laranja', então João instancia uma estrutura neural isomórfica a essa proposição. Ou seja, João instancia uma estrutura simbólica constituída de unidades de significado com os conteúdos 'alienígenas', 'bolo de laranja' etc. e, além disso, essas unidades estão arranjadas de maneira correspondente ao modo como os conceitos estão dispostos naquela proposição. O atomismo semântico é a tese de que os conceitos que João utiliza para formular essa proposição mental são primitivos ou constituídos de conceitos primitivos através de um mecanismo formal de composição simbólica. Por exemplo, 'bolo de laranja' não é um conceito primitivo, pois é formado através da combinação de ideias mais simples, 'bolo' e 'laranja (fruta)'.

O atomismo semântico se contrapõe a teses holísticas do significado (Jackman (2020)), de acordo com as quais o conteúdo semântico de um termo (e.g. 'bolo de laranja') é individuado segundo o lugar que ele ocupa na totalidade de uma rede de significados (e.g. léxico do

português)[52]. Suponha que o nosso repertório de representações mentais seja algo como um dicionário. Em um dicionário, os termos são todos interdefinidos. Isso significa que qualquer mudança no significado de um dos termos afeta toda a rede através de um efeito dominó. Se o significado de um termo T muda, isso afetará o significado de todos os outros que estão diretamente relacionados a T e, consequentemente, todos aqueles que estiverem diretamente relacionados aqueles que estão diretamente relacionados a T, e assim por diante, até que toda a rede seja alterada.

Não entraremos em detalhes, mas o holismo semântico parece ter implicações destrutivas para a psicologia de senso comum. Uma vez que o significado de um termo é individuado segundo o lugar que ele ocupa na rede de significados, disso se segue que duas pessoas não podem instanciar a mesma atitude proposicional, pois duas pessoas nunca terão exatamente a mesma rede de representações mentais. De acordo com o atomismo semântico, por outro lado, é completamente possível que duas pessoas compartilhem os mesmos primitivos representacionais. É inteiramente possível, portanto, que duas pessoas instanciem a mesma atitude proposicional.

Mas essa não é a única motivação da TCM em favor do atomismo. Uma das principais razões que Fodor oferece em defesa dessa tese é a capacidade que ela tem de explicar a *sistematicidade* e *produtividade* da linguagem e do pensamento. Considere as línguas naturais. Qualquer falante nativo do português é capaz de entender o excerto a seguir:

> "Não há muitas pessoas que saibam que, em 1931, Adolf Hitler foi aos EUA, visitou vários pontos de interesse, teve em Keokuk, Iowa, um caso amoroso com uma senhora de nome Maxine, experimentou *peyote* (o que o fez ter alucinações com hordas de rãs e sapos que calçavam botinhas vermelhas e cantavam o *Horst Wessel Lied*), infiltrou-se numa fábrica de munições perto de Detroit, encontrou-se secretamente com o vice-presidente Curtis para tratar de futuros compromissos comerciais relativos às peles de foca e inventou o abre-latas eléctrico."

Ao não ser que o leitor esteja familiarizado com o livro de Lycan (2000), *Filosofia da Linguagem*, provavelmente nunca leu o texto acima. Até 2000, ano em que esse livro foi escrito e publicado, essa ordem específica de palavras não havia sido escrita. Não obstante, não há dificuldade alguma por parte dos falantes nativos do português em entendê-la integralmente. Isso ilustra a sistematicidade e a produtividade da língua portuguesa. A *produtividade* é a capacidade

[52] No Capítulo 2 (§2.5.1) vimos uma vertente holística do significado, a Conceptual Role Semantics, que individua o conteúdo semântico de uma atitude proposicional em termos do papel funcional que ela ocupa na dinâmica cognitiva do indivíduo.

do sistema de produzir infinitas sentenças novas, com significados nunca antes formulados. A *sistematicidade* é o conjunto de regras do sistema que estabelece como formular esses novos significados a partir da manipulação dos elementos significativos (palavras). A língua portuguesa possui uma *semântica combinatória*, o que significa que novos significados são obtidos a partir de diferentes combinações de elementos lexicais, segundo regras gramaticais específicas.

A ideia de Fodor (1975, 1987) é que a sistematicidade e produtividade das línguas naturais derivam da sistematicidade e produtividade do pensamento. Essa não é uma conclusão absurda, já que a língua natural é, em certo sentido, uma expressão do pensamento. Assumindo que nossos pensamentos possuem uma semântica combinatória, portanto, podemos entender como eles (e as línguas naturais) são sistemáticos e produtivos. Vale observar que uma semântica combinatória pressupõe o atomismo, já que a combinação nada mais é do que a união ou encadeamento de elementos primitivos formando estruturas moleculares mais complexas e, assim, gerando novos significados.

3.2.2 INDIVIDUAÇÃO DE ESTADOS MENTAIS

Outro ponto importante acerca do posicionamento da TCM com relação às teorias semânticas contemporâneas diz respeito a como individuamos nossos estados mentais. Desde Frege (2009), fazemos a distinção entre o *sentido* de um termo ou proposição e o seu *referente* no mundo. Frege nos mostrou que dois estados mentais podem ter conteúdos (sentidos) diferentes, embora se refiram ao mesmo objeto. Por exemplo, o pensamento de que a Torre Eiffel fica em Paris e o pensamento de que a maior estrutura de aço da França fica em Paris são diferentes, embora 'Torre Eiffel' e 'maior estrutura de aço da França' compartilhem o mesmo referente no mundo. São dois *modos de apresentação* de um mesmo referente.

Podemos fazer uma analogia com os sentidos. Quando duas pessoas observam uma árvore que está a certa distância, terão estados mentais cujo referente é o mesmo, mas a árvore se apresentará a eles de maneiras distintas, pelo fato de que eles ocupam posições diferentes em relação a ela. Nesse sentido, parece razoável individuar os estados mentais dos dois como sendo estados mentais diferentes, porque consistem em experiências sensoriais qualitativamente diferentes.

O mesmo poderia ser dito do pensamento de que a Torre Eiffel fica em Paris e do pensamento de que a maior estrutura de aço da França fica em Paris. Eles possuem valores cognitivos diferentes. Uma maneira de individuar estados mentais é levando em conta apenas esse valor cognitivo ou modo de apresentação dos pensamentos. De acordo com essa visão,

chamada de *internalismo semântico*, um cérebro numa cuba que vive em uma realidade virtual pode ter os mesmos estados mentais que uma pessoa que vive fora da realidade virtual. O que importa é se eles instanciam estados mentais qualitativamente idênticos. Outra maneira de caracterizar o internalismo semântico é dizendo que a individuação de um estado mental precisa levar em conta apenas seus aspectos próprios, locais ou internos, em oposição a aspectos etiológicos e relações com o ambiente externo.

Entretanto, alguns problemas surgem quando pensamos na relação entre o conteúdo de nossos pensamentos e os objetos no mundo aos quais eles se referem. A questão crucial aqui é "Em virtude de que meu pensamento se refere a um objeto específico no mundo?". O conteúdo de meu estado mental parece cumprir um papel importante nesse processo. 'A estrutura de aço mais alta da França' se refere à Torre Eiffel, alguém poderia dizer, porque essa descrição se aplica unicamente a ela. Essa pessoa poderia continuar: 'Na verdade, o termo 'Torre Eiffel' é sinônimo de um conjunto de descrições – 'estrutura de aço mais alta da França', 'possui uma base com quatro pés' etc. – que unicamente selecionam o mesmo objeto no mundo, e é assim que nossos pensamentos fazem referência a uma coisa e não a outra'. De acordo com essa tese, conhecida como *descritivismo*, o sentido de termos próprios, como 'Torre Eiffel', 'Albert Eistein' e 'Salma Hayek' equivale a um conjunto de descrições que se aplica a um único objeto, e é por isso que pensamentos que envolvem esses termos próprios são sobre esses objetos. Seguindo o espírito do internalismo semântico, o descritivismo tenta explicar como nossos pensamentos se referem a objetos externos levando em conta apenas aspectos internos de nossos estados mentais.

Porém, Kripke (1972) nos mostra, quase com a força de uma demonstração matemática, que termos próprios não podem ser sinônimos de descrições, porque termos próprios são *designadores rígidos*, e descrições são *designadores não-rígidos*. Um designador rígido é um termo que se aplica sempre a um mesmo objeto. Um designador não-rígido é uma expressão que se aplica a diferentes objetos em diferentes contextos. 'Torre Eiffel' se aplica sempre à Torre Eiffel, não importa o que aconteça. 'A estrutura de aço mais alta da França', porém, se aplica à Torre Eiffel de maneira contingente, pois se uma outra torre de aço fosse construída e fosse mais alta que a Torre Eiffel, essa expressão se aplicaria a ela.

Um movimento contrário ao internalismo semântico ganha força, sustentando a ideia de que a referência ocorre quando há uma conexão causal entre nossos estados mentais e os objetos aos quais eles se referem. Putnam (1975c) avança um experimento mental que tenta provocar a intuição de que descrições mentais e aspectos qualitativos de nossos estados psicológicos são insuficientes para selecionar um único objeto no mundo.

De acordo com o experimento mental, há um outro planeta do outro lado do universo que é praticamente idêntico ao nosso, a Terra Gêmea. A única diferença é que não há H_2O na Terra Gêmea, mas uma substância química, XYZ, cujas propriedades são indistinguíveis das propriedades da H_2O. Além disso, todas os terráqueos possuem uma contraparte na Terra Gêmea. Oscar, um terráqueo, possui um "irmão gêmeo" da Terra Gêmea, Toscar, e ambos possuem exatamente a mesma história de vida, tiveram as mesmas experiências e instanciam exatamente os mesmos estados psicológicos. O argumento de Putnam é o de que quando Oscar pensa sobre água, ele pensa sobre H_2O, enquanto que quando Toscar pensa sobre água, ele pensa sobre XYZ. Oscar e Toscar possuem estados mentais qualitativamente idênticos, e atribuem exatamente a mesma descrição à água — líquido incolor, inodoro, insípido, encontrado nos lagos e que usamos para beber, lavar roupa, cozinhar etc. No entanto, quando Oscar afirma 'Água evapora a 100°C', essa afirmação é verdadeira se, e somente se, H_2O evaporar a 100°C. Quando Toscar afirma 'Água evapora a 100°C', essa afirmação é verdadeira se, e somente se, XYZ evaporar a 100°C.

Como é possível que dois estados mentais com o mesmo conteúdo se refiram a objetos diferentes no mundo? Ou admitimos que o conteúdo (qualitativo ou descritivo) de nossos estados psicológicos são insuficientes para selecionar objetos como referência, ou admitimos que o conteúdo de nossos estados mentais não é esgotado ou determinado unicamente por essas características qualitativas, descritivas e internas ao próprio estado mental. O *externalismo semântico* é a tese de que parte do conteúdo de nossos estados mentais é determinado pela relação causal que eles possuem com o ambiente externo que os gerou em primeiro lugar. Filósofos começam a fazer uma distinção entre dois tipos de conteúdo. Oscar e Toscar compartilham o mesmo conteúdo estrito (*narrow content*), mas não o mesmo conteúdo amplo (*wide content*).

Isso é importante porque, como iniciamos a seção, há uma grande questão a respeito de como estados mentais devem ser individuados. É fácil de ver que a crença de que a Torre Eiffel fica em Paris possui conteúdo distinto da crença de que a maior estrutura de aço da França fica em Paris, e parece intuitivamente correto, portanto, individuá-las como atitudes proposicionais distintas. O que é difícil de aceitar é que as atitudes proposicionais de Oscar e Toscar a respeito de água sejam distintas, pois não apenas os estados mentais de ambos são qualitativamente idênticos, mas possuem propriedades causais similares (Fodor (1987)).

Mas, como veremos na próxima seção, a relação semântica que os estados internos de um organismo possuem com o ambiente externo é o que nos permite individuar as computações ele está implementando quando essa individuação é ambígua, i.e., quando há mais de uma

interpretação computacional possível a respeito do processo físico ocorrendo no hardware. Desse modo, a questão a respeito de como propriedades semânticas são relevantes na individuação de estados mentais é importante para nos ajudar a responder à questão de quais computações um sistema implementa.

John Searle argumenta que computações concretas não existem porque computações consistem na atribuição aleatória de propriedades sintáticas a um sistema. Argumentaremos na próxima seção que essa atribuição de propriedades sintáticas não é aleatória, pois ela respeita a relação semântica não arbitrária que existe entre os estados internos do sistema e o ambiente no qual ele está inserido.

3.3 PROPRIEDADES SINTÁTICAS

John Searle argumenta que o grande problema do computacionalismo é que 'sintaxe' é uma noção essencialmente relativa a um observador. A conclusão seria, é claro, que qualquer processo computacional existiria apenas relativamente a alguém que fez a atribuição sintática arbitrária em primeiro lugar.

> But these further restrictions on the definition of computation are no help in the present discussion because *the really deep problem is that syntax is essentially an observer-relative notion. The multiple realizability of computationally equivalent processes in different physical media is not just a sign that the processes are abstract, but that they are not intrinsic to the system at all. They depend on an interpretation from outside* (SEARLE 1992, p. 209, grifo do autor).

Mas Searle vê a atribuição de propriedades sintáticas a sistemas concretos como uma atividade completamente arbitrária porque ele rejeita a possibilidade de que os estados internos do sistema (no caso, do cérebro) possuam conteúdo semântico intrínseco (no nível inconsciente). Ele utiliza o fato de que podemos construir, a partir de qualquer substrato físico, sistemas organizados com propriedades formais atribuídas como uma evidência para inferir que nenhum sistema possui propriedades sintáticas intrínsecas. É verdade que níveis voltaicos, polaridades magnéticas e alavancas não possuem sintaxe essencialmente, mas isso se deve ao fato de que esses substratos *não possuem uma semântica intrínseca*. Uma vez que o conteúdo dos estados internos desses sistemas é relativo ao observador, a sintaxe também o é. Mas, suponha por um instante que os estados internos de uma máquina carreguem conteúdo semântico intrínseco. Nesse caso, não poderíamos projetar nela características formais da maneira que quiséssemos, porque a

caracterização sintática desse sistema deve respeitar as relações semânticas que já existem. Para ilustrar essa ideia, considere o seguinte exemplo fictício, adaptado de Oron Shagrir (2006) — *Why we view the brain as a computer*.

Suponha que o cérebro do sapo possua um sistema detector de moscas pretas chamado S. Quando uma mosca preta aparece em seu campo visual, S é imediatamente alertado e emite um impulso nervoso que coordenada os movimentos da língua do sapo, a qual é rapidamente atirada em direção ao objeto detectado. Se uma mosca de outra cor aparece — digamos, vermelha —, S não é acionado. Analogamente, se um objeto preto que não é uma mosca se tornar visível para o sapo, S permanece inativo.

Como esse sistema funciona? S possui duas entradas, as quais recebem sinais de canais diferentes, e apenas uma saída. O primeiro canal emite um sinal elétrico de 50-100mV sempre que uma mosca aparece em cena, e o segundo emite de 50-100mV sempre que um objeto preto se torna visível (assuma que os canais são sempre sensíveis ao mesmo objeto). Se não há nenhuma mosca ou objeto preto à vista, os canais emitem um sinal elétrico de 0-50mV. S, além de receber, também emite sinais elétricos nas duas faixas voltaicas; se S recebe dois sinais de 50-100mV, responde com uma saída de 50-100mV (que gera o movimento de captura com a língua). Em qualquer outra situação, S responde com uma saída de 0-50mV, a qual não excita nenhuma resposta motora.

Canal 1 (mV)	Canal 2 (mV)	Saída de S (mV)
50-100	50-100	50-100
50-100	0-50	0-50
0-50	50-100	0-50
0-50	0-50	0-50

Tabela 2.1 - Os canais 1 e 2 servem de entrada para o sistema S. A saída de S é conforme a tabela. A interpretação computacional natural do comportamento do sistema consiste na implementação da função AND.

A interpretação computacional de S é óbvia: S funciona como uma porta AND, e as faixas voltaicas de 0-50mV e 50-100mV devem ser interpretadas, respectivamente, como 0's e 1's. Nós descrevemos S como um sistema detector de moscas pretas porque ele une, a partir de um processo computacional, informações sobre moscas e informações sobre coisas pretas, fornecendo um sinal que representa a intersecção dos dois conjuntos. Dado que um canal representa a presença de moscas e o outro canal representa a presença de objetos pretos, e dada a função AND implementada, a saída de S entre 50-100mV deve representar moscas (&) pretas.

Entretanto, suponha que as bandas voltaicas sejam um pouco mais refinadas do que foi descrito anteriormente. Assuma que S emite 50-100mV quando recebe dois sinais de 50-100mV, mas que ele emite 0-25mV sempre que recebe dois sinais de 0-25mV, e 25-50mV em qualquer outra situação. Nesse caso, podemos interpretar sinais de 0-25mV como 0's e sinais de 25-100mV como 1's. Sob essa atribuição sintática, S implementa uma porta OR, o que significa que o mesmo sistema físico possui ao menos duas interpretações computacionais possíveis.

Canal 1 (mV)	Canal 2 (mV)	Saída de S (mV)
50-100	50-100	50-100
50-100	25-50	25-50
50-100	0-25	25-50
25-50	50-100	25-50
25-50	25-50	25-50
25-50	0-25	25-50
0-25	50-100	25-50
0-25	25-50	25-50
0-25	0-25	0-25

Tabela 3.2 - A Tabela 2 é apenas um refinamento da Tabela 1, de modo que o comportamento do sistema é apresentado em maior resolução voltaica. S continua sendo idêntico ao caso anterior.

Ecoamos o argumento de Shagrir a favor da ideia de que, embora os estados físicos e as transições de estados físicos de S sejam compatíveis com a interpretação formal correspondente à porta OR, a estrutura computacional de S corresponde à porta AND, e esse fato é determinado, ao menos em parte, pelo conteúdo informacional que seus estados internos carregam. Se interpretarmos os sinais acima de 50mV como contendo certa informação (sobre moscas, coisas pretas e moscas pretas), e entendendo como as operações internas de S favorecem a sobrevivência do sapo, é possível apreciar o seu valor adaptativo e, portanto, entender por que S opera da maneira que opera — como uma porta AND. Por outro lado, não há nenhuma boa razão para interpretarmos o comportamento de S como implementando a função OR.

A primeira conclusão é a de que um simples isomorfismo estrutural entre um algoritmo e um sistema complexo qualquer não é razão suficiente para dizer que esse sistema implementa esse algoritmo. Isso, no entanto, Searle está disposto a aceitar. Sua maior preocupação é com a tese de que podem haver boas razões para atribuir propriedades sintáticas que sejam intrínsecas ao sistema. Baseado em que? O argumento de Searle é o de que atribuições sintáticas são

completamente arbitrárias. Podemos interpretar polaridades magnéticas de um computador da maneira que quisermos. Isso, segundo o autor, revela não apenas que as propriedades sintáticas de polaridades magnéticas não são apenas abstrações, mas evidências de que as propriedades sintáticas não pertencem ao sistema de modo algum. São qualidades relativas a um observador que atribui predicados sintáticos da maneira que for conveniente. A segunda conclusão que tiramos do exemplo acima, e que é a mais importante para nós, é que é falsa a tese de que atribuições sintáticas são completamente arbitrárias. O conteúdo semântico dos estados internos do sistema restringe sua estrutura computacional. A arbitrariedade da interpretação sintática de sistemas computacionais construídos por nós é apenas um sinal de que esses sistemas não possuem uma semântica intrínseca. Mas, assumindo que estados cerebrais possuem conteúdo semântico intrínseco (até mesmo nos níveis inconscientes), podemos ver como esse conteúdo exerce um papel na individuação dos processos formais.

Searle às vezes parece sugerir que a TCM assume um nível *ontológico* do sistema caracterizado por propriedades sintáticas:

> The implemented program has no causal powers other than those of the implementing medium because the program has no real existence, no ontology, beyond that of the implementing medium. Physically speaking, there is no such thing as a separate "program level" (SEARLE 1992, p. 215)

O autor argumenta que explicações a respeito da cognição precisam ser explicações *causais*. Como a sintaxe não possui poderes causais além daqueles já possuídos pelo veículo implementacional, explicações sintáticas seriam no mínimo supérfluas. Na verdade, de acordo com o autor, essas explicações seriam falsas porque sequer existe um nível ontológico de propriedades formais. Mas Searle interpreta mal a TCM. A Teoria Computacional da Mente não afirma que o nível sintático é um nível ontológico à parte, mas um nível de *análise* distinto. Temos que ter em mente o conceito de níveis de explicação apresentado no Capítulo 1. Níveis de explicação distintos usam taxonomias distintas e, por essa razão, fazem generalizações diferentes. Com generalizações diferentes, abre-se a possibilidade de se fazer explicações e previsões que não são possíveis de se fazer em outros níveis de abstração. A caracterização formal das operações do cérebro não implica num nível ontológico diferente do nível neural. Ao contrário, como vimos no primeiro capítulo, o nível formal é instanciado pelo nível físico. O ponto é que essa caracterização formal se abstrai de particularidades neurofisiológicas porque essa caracterização formal pode se aplicar a sistemas com particularidades neurofisiológicas distintas. Assim, poderíamos elaborar explicações semelhantes e fazer previsões análogas que se

aplicariam a diferentes sistemas em virtude de compartilharem as mesmas propriedades formais (ver Apêndice).

Por que dois sistemas compartilhariam as mesmas propriedades formais? Bem, se olharmos para o sistema nervoso central de uma perspectiva informacional, é plausível que organismos diferentes tenham encontrado soluções parecidas para os mesmos problemas evolutivos. Se encontramos homogeneidades fisiológicas em diferentes espécies animais, podemos encontrar homogeneidades cognitivas também. Essa hipótese respeita o princípio evolutivo mais básico da biologia. O modo como nossos órgãos evoluíram é explicado pela função adaptativa que eles cumpriram na sobrevivência de nossos ancestrais. Se esses órgãos nos ajudaram a sobreviver, também nos ajudaram (indiretamente) a gerar cópias genéticas de nós mesmos. Famílias de animais diferentes tiveram a oportunidade de desenvolver características parecidas, por mero acaso genético. Mas, pelo fato de que essas características apresentaram valor adaptativo, temos famílias de animais de linhagens filogenéticas distintas com soluções evolutivas convergentes[53]. Aplique essa história ao cérebro. O modo como um organismo interpreta e processa informações a respeito do ambiente externo, de outros animais e de si mesmo tem um valor adaptativo decisivo na vida desse organismo. Espécies diferentes podem ter encontrado soluções parecidas no modo como informações a respeito do ambiente são apreendidas, representadas e tratadas para uso geral. Não surpreendentemente, encontramos convergências cognitivas em animais de linhagens evolutivas completamente distintas, como o caranguejo, o gato, e seres humanos (Lindsay & Norman 1977)[54].

Para resumir o que foi dito até agora, portanto, acreditamos que Searle interpreta de maneira equivocada as assunções teóricas da TCM. Em particular, consideramos que ele falha em entender papel teórico da dimensão sintática de explicação dos processos formais. A Teoria Computacional da Mente não atribui uma ontologia às propriedades sintáticas diferente da ontologia dos veículos físicos que implementam os processos cognitivos. Ao nosso ver, as propriedades sintáticas são abstrações que visam capturar generalizações impossíveis de serem observadas se atentarmos unicamente ao nível do hardware do sistema.

3.4 A VISÃO MECANICISTA DE COMPUTAÇÃO CONCRETA

[53] O inverso também é possível. Em geral, grupos da mesma espécie que são separados geograficamente (e, portanto, são expostos a ambientes diferentes) tendem, a longo prazo, a desenvolver características fenotípicas diferentes, dando origem a espécies distintas. São casos de divergência adaptativa.
[54] Ver Apêndice.

Há outras correntes teóricas alternativas à abordagem semântica que também tentam esclarecer a noção de computação concreta (ver Piccinini & Maley (2021)). A Simple Mapping Account é uma delas, mas, como vimos, essa alternativa incorre no pancomputacionalismo. As principais correntes que disputam com a visão semântica são a visão sintática e a visão mecanicista de computação (idem). Apresentaremos brevemente uma dessas opções, viz., a abordagem mecanicista.

Piccinini (2015) é um dos maiores defensores da visão mecanicista de computação, segundo a qual um sistema computacional é individuado de acordo com as funções executadas por suas partes internas. Mais precisamente, um sistema computacional é um mecanismo funcional tal que uma de suas funções é executar computações. De acordo com essa corrente, não é necessário que um sistema computacional opere sobre estruturas representacionais, i.e., não é necessário que os veículos manipulados carreguem conteúdo semântico – embora frequentemente esse seja o caso. Porém, a visão mecanicista impõe outras restrições sobre a definição de computação concreta, de modo que o pancomputacionalismo seja evitado. Algumas dessas restrições são:

- Sistemas físicos que não são *mecanismos funcionais* não podem ser sistemas computacionais;

- Mecanismos que não possuem a função de manipular *veículos independentes do meio* não são candidatos a sistemas computacionais;

- Mecanismos que manipulam veículos independentes do meio, mas que falham em seguir *regras apropriadas*, não são candidatos a sistemas computacionais.

A classe de objetos que se enquadram nesses requisitos é a classe de objetos que constituem sistemas computacionais concretos (Piccinini & Maley (2021)). Os termos destacados acima serão explicados a seguir.

Em geral, mecanismos funcionais são definidos conforme o papel ou propósito que cumprem dentro de um sistema. Desse modo, uma ratoeira pode ser caracterizada como o objeto que captura roedores e o coração pode ser caracterizado como o órgão responsável por bombear sangue para outros órgãos do corpo. Note que em nenhuma dessas caracterizações está especificado o substrato físico que ratoeiras ou corações precisam ter. Uma ratoeira pode ser feita de plástico ou de metal, e um coração pode ser feito de tecido biológico ou de material biônico.

Entretanto, há uma diferença entre ratoeiras e corações, qual seja, a de que uma é projetada para cumprir essa função, o outro, não – ao menos, não no mesmo sentido. Hoje,

sabemos que cada parte do corpo é fruto de um processo acumulativo de seleção, sem que tenha havido intenção inteligente por trás desse processo. Como Searle (1992) argumenta, afirmar que o coração possui a função de bombear sangue é apenas uma maneira de caracterizar esse órgão com relação aos nossos próprios interesses. Afirmar que o coração possui a função de bombear sangue é dizer que ele foi intencionalmente projetado para isso, que ele *deve* fazer isso. Entretanto, argumenta o autor, essa ideia é completamente ortogonal ao que foi proposto pela teoria da evolução. Por exemplo, se as propriedades sonoras do coração fossem de nosso interesse, teríamos classificações de doenças cardíacas diferentes das que temos atualmente (Searle (1992)).

O mesmo argumento vale para o cérebro. Parece inconsistente com a teoria da evolução que o cérebro seja um mecanismo com funções teleológicas, com subsistemas que possuem propósitos específicos, uma vez que a noção de *propósito* parece pressupor a noção de *planejamento intencional*. Piccinini, porém, argumenta que há fatos não teleológicos no mundo que determinam características teleológicas em organismos. De acordo com o autor, um traço (ou um componente, atividade ou propriedade) de um organismo é funcional se ele for uma contribuição estável para os objetivos desse organismo (Piccinini (2015)). No caso de nossos órgãos, suas funções são determinadas pelo valor adaptativo que tiveram em nosso histórico evolutivo. É por isso que o batimento cardíaco possui função, mas o som das batidas, não.

Mas, ser um mecanismo funcional não é suficiente para ser um sistema computacional. Piccinini argumenta que o cérebro é um sistema computacional porque ele manipula veículos independentes do meio de acordo com regras apropriadas. Um veículo é um componente espaço-temporal do sistema que corresponde a uma variável ou um valor dessa variável – i.e., um componente que *funciona* como uma variável ou valor da variável. As regras são relações de mapeamento entre inputs e outputs do sistema. O veículo será independente do meio se as regras que caracterizam os processos internos do sistema forem sensíveis apenas às dimensões físicas dos veículos que são relevantes para as computações. Em geral, os veículos precisam ser multiplamente realizáveis.

Um dos argumentos de Piccinini a favor da visão mecanicista é que, diferentemente de outras abordagens, o mecanicismo teleológico consegue acomodar falhas computacionais. Suponha que um sistema computacional possui a função de computar, a partir do input a, $f(a)$. Se em determinada situação esse sistema produzir um resultado diferente do usual, em virtude de ruídos ou interferências, a visão semântica teria que dizer apenas que o sistema computou uma função diferente, digamos, $g(a)$. No caso do mecanicismo, há um componente normativo

que determina que o sistema *deveria* ter computado $f(a)$. Portanto, argumenta o autor, o mecanicismo consegue diferenciar ou individuar casos em que o sistema falha.

Entretanto, a abordagem mecanicista também sofre críticas. Boccardi (2009), por exemplo, argumenta que o funcionalismo é insuficiente para individuar sistemas computacionais, e que seria necessário apelar para outras propriedades (semânticas ou físicas) para se evitar uma caracterização circular dos componentes do sistema. Essa discussão, porém, está além do escopo deste trabalho.

4. ALTERNATIVAS TEÓRICAS AO COMPUTACIONALISMO CLÁSSICO

Por algum tempo, o computacionalismo clássico foi a teoria dominante nas ciências cognitivas e, embora ainda haja diversos programas de pesquisa erigidos sobre essa base teórica, o tema é muito mais controverso hoje em dia. Neste último capítulo, faremos uma breve exposição a respeito de duas importantes abordagens adversárias ao computacionalismo clássico, a saber, o modelo de processamento distribuído paralelo e a teoria da cognição incorporada e situada.

A primeira seção apresenta o modelo de processamento distribuído paralelo (PDP), que aborda nossas funções cognitivas de maneira distinta da abordagem computacional. Ao contrário da metodologia top-down empregada por Marr, por exemplo, defensores dos modelos de PDP defendem uma coevolução entre os diferentes níveis de investigação do cérebro, de modo que cada nível oferece suas próprias contribuições e insights de maneira independente, podendo, inclusive, corrigir e contribuir com as investigações dos níveis mais altos de análise dos processos mentais. Além disso, esses modelos são biologicamente inspirados, i.e., eles imitam redes neurais e tentam executar tarefas cognitivas por meio de conexões entre unidades de processamento que imitam células nervosas.

Na segunda seção, vemos uma outra abordagem, chamada de cognição incorporada e situada (CIS). De acordo com essa corrente teórica, as ciências cognitivas centraram a atenção exclusivamente no cérebro, quando na verdade nossos processos cognitivos são constituídos por todo o corpo, levando em conta o ambiente e contexto sócio cultural nos quais o organismo está inserido. Parte da motivação dos teóricos da CIS se baseia na dificuldade que as outras abordagens têm – sobretudo, o computacionalismo – em esclarecer uma série de problemas relacionados às nossas habilidades cognitivas e manifestações comportamentais. Eles acreditam que esses problemas são solucionáveis se levarmos em conta o papel do corpo na cognição.

4.1 PROCESSAMENTO DISTRIBUÍDO PARALELO

Como vimos nos capítulos anteriores, a Teoria Computacional da Mente defende a ideia de que uma compreensão global dos processos cerebrais envolve uma série de níveis de caracterização, os quais estão associados a diferentes níveis de organização do cérebro. No caso das ciências cognitivas, a estratificação disciplinar imposta por essa multiplicidade de escalas (da micro até a macro) é tão ineliminável quanto intrigante. Diferentes disciplinas requerem diferentes ferramentas teóricas e tecnológicas de investigação, incluindo a própria estrutura

conceitual sobre a qual elas se apoiam para formular hipóteses, fazer previsões e oferecer explicações. Tanto o neurocientista celular, na tentativa de identificar e compreender as funções das células nervosas em nível microscópico, quanto o psicólogo cognitivo, na empreitada de descrever os mecanismos mentais que causam uma ilusão de óptica, por exemplo, estão, em um sentido importante, estudando o funcionamento do cérebro – muito embora o neurocientista e o psicólogo não compartilhem o mesmo vocabulário, as mesmas ferramentas ou o mesmo laboratório.

Mas é natural esperar que existam pontos de contato entre esses diferentes níveis explicativos. Na ciência, é comum que teorias mais fundamentais expliquem e legitimem princípios teóricos menos fundamentais. Por exemplo, a lei de que a água entra em ebulição a 100°C (em condições normais de temperatura e pressão) possui uma explicação teórica que recorre a elementos mais fundamentais da realidade. Essa lei é explicada quando vemos a água não como um material bruto do universo, mas como um conjunto de componentes, moléculas, que possuem propriedades tais que, sob certas condições, se comportam como um líquido e, sob outras, como um gás. Parece razoável esperar que esse tipo de explicação vertical também ocorra nas ciências cognitivas. Dessa perspectiva, princípios psicológicos poderiam muito bem ser justificados recorrendo-se ao corpo teórico das ciências do cérebro (ver Apêndice).

Seguindo esse raciocínio, os teóricos da TCM e, notavelmente, David Marr, sugeriram um modelo investigativo em que se reconhece não apenas os pontos de contato entre os diferentes níveis investigativos das ciências do cérebro e da mente, mas *como* esses pontos de contato serão estabelecidos.

Basicamente, o modelo segue o método da engenharia reversa. O cérebro é visto como uma poderosa máquina processadora de informações, capaz de executar complexas funções. A primeira coisa a se fazer é caracterizar essas funções, i.e., identificar os inputs e outputs envolvidos em cada uma delas. Por exemplo, no caso de Marr (1982), o autor está interessado em saber como o cérebro é capaz de produzir representações tridimensionais a partir de estímulos luminosos que incidem sobre a superfície da retina. Imagine que a luz refletida em um sólido euclidiano é projetada numa parede. A parede, nesse caso, é análoga ao olho, e o que o cérebro precisa fazer é reverter computacionalmente essa projeção, ou seja, inferir a forma do objeto tridimensional a partir de sua projeção bidimensional (Burge (2010)). É um problema matemático. A ideia é a de que o sistema visual resolve essa tarefa decompondo-a em funções menores — identificando bordas, arestas, ângulos e muitas outras coisas. Assim, a segunda coisa a se fazer é descobrir quais *algoritmos* são implementados pelo cérebro, tal que as informações das matrizes de luz bidimensionais projetadas nos olhos sejam transformadas em informações

sobre bordas, arestas, ângulos etc. e, posteriormente, reunidas em uma representação tridimensional. Tendo encontrado uma solução computacional psicologicamente plausível, resta investigar os mecanismos neurais empregados pelo cérebro para implementar esses algoritmos. Em resumo, o roteiro de pesquisa é 1) caracterizar a função cognitiva que requer explicação, 2) encontrar os algoritmos que solucionam essa função e 3) investigar como o cérebro implementa esses algoritmos neurofisiologicamente (ver §1.4.1).

Esse tipo de explicação ficou conhecido como explicação de cima para baixo (*top-down explanation*), pois ela se inicia identificando os processos cognitivos de larga escala, em direção aos mecanismos neurais de baixa escala. Esse sentido investigativo prioriza o topo da hierarquia de níveis, pois a função de um nível investigativo é a de explicar como e por que ocorrem os princípios e generalizações do nível acima. Primeiro, o psicólogo cognitivo se encarrega de caracterizar quais funções o cérebro executa. Depois, modelos computacionais são considerados para explicar como essas funções poderiam ser levadas a cabo. Por último, neurocientistas mapeariam o cérebro em busca dos recursos neurais que implementariam fisicamente os programas.

O problema com esse modelo é o de que não parece haver espaço para uma mudança no modo como entendemos a psicologia que emerja de baixo para cima. A psicologia é vista como um nível de explicação privilegiado e autônomo, ao contrário dos outros níveis explicativos, cujas funções são, em última instância, esclarecer como o cérebro emprega mecanismos neurais que garantem a veracidade das generalizações feitas pela psicologia. De fato, a maioria dos autores reconhece o caráter "privilegiado" da psicologia (Bermúdez (2005)). David Braddon-Mitchell e Frank Jackson, por exemplo, argumentam que apenas por um milagre os princípios da psicologia de senso comum seriam falsos, dado o poder explicativo e preditivo que eles nos fornecem:

> Trees and planets behave in relatively regular ways [...] By contrast, human beings move in a quite bewildering variety of ways. Nevertheless, we often succeed in predicting what they will do. How do we do this? By treating them as subjects with mental states. By observing what they do and say, we arrive at views about what they are thinking, what they desire and closely associated views about their characters, mental capacities and in general about their psychological profiles. We then, in terms of these profiles, predict what they will do [...] Think of what is involved in playing a game of tennis, crossing a road at traffic lights or organizing a conference. The antecedent probability that Jones will move her body in such a way that the ball will land where you have most trouble retrieving it, or that drivers will move their bodies in such a way that their cars will stop when the light turns red, or that a number of human bodies will move from various corners of the globe to end up at the same time in one conference centre, is fantastically small. Yet we make such predictions successfully all the time...The fact that we can make the predictions shows that we have cottoned on to the crucial regularities – otherwise our predictive capacities would be a miracle. They show that we have an implicit mastery of a

> detailed, complex scheme that interconnects inputs, outputs and mental states. (BRADDON-MITCHELL & JACKSON, 1996, p. 56-57)

Jerry Fodor, de maneira mais dramática, afirma que se a psicologia de crenças e desejos for falsa, esse é o fim do mundo.

> Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying..., if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world. (FODOR, 1992, p. 156)

Mas os críticos do computacionalismo estão abertos a outras possibilidades. Eles acreditam que o avanço da neurociência e da modelagem computacional podem revelar insights que mudem completamente o modo como entendemos a mente humana. Os defensores do modelo de processamento distribuído paralelo (PDP) argumentam que a abordagem explicativa de cima para baixo deve ser suplantada por uma abordagem co-evolutiva de investigação, na qual há uma troca recíproca entre os diferentes níveis de explicação (Bermúdez (2005)). De acordo com essa visão, é claro que o neurocientista deve estar atento às descobertas feitas pelo psicólogo cognitivo. Mas é inteiramente possível, ao mesmo tempo, que o psicólogo revise suas assunções à luz de descobertas feitas pela neurociência, mesmo que essas assunções sejam basilares em sua disciplina.

Essa é uma diferença importantíssima entre a TCM e o modelo PDP. Os defensores da TCM argumentam que os padrões sublinhados pela psicologia de crenças e desejos devem possuir padrões correspondentes nos níveis mais fundamentais de caracterização das atividades do cérebro. Dito de outro modo, se no nível psicológico emergem padrões mentais específicos, é porque eles são meras manifestações de padrões isomórficos que ocorrem no nível neural. Afinal, qual outra maneira de explicar as regularidades detectadas pela psicologia? No entanto, como veremos, a neurociência computacional guarda uma compreensão completamente diferente de nossa arquitetura cognitiva, de acordo com a qual os padrões mentais capturados pela psicologia podem emergir de operações neurais que não são, necessariamente, estruturalmente isomórficas a eles.

Para entendermos esse programa de pesquisa, precisamos conhecer um pouco o modelo computacional utilizado para estudar nossas funções psicológicas.

4.1.1 REDES NEURAIS ARTIFICIAIS

Embora tenhamos visto, nas últimas décadas, um enorme avanço científico nas técnicas de escaneamento cerebral, o fato é que essas ferramentas ainda são insuficientes para iluminar as atividades de média escala do cérebro (Bermúdez (2014)). Técnicas de mapeamento, como imagem por ressonância magnética funcional (fMRI) e eletroencefalograma (EEG), são muito úteis para detectar atividade em grandes populações de neurônios. É possível mapear o fluxo sanguíneo no cérebro (fMRI) ou as atividades elétricas geradas pelos disparos neurais (EEG) e, com base nisso, determinar quais regiões cerebrais estão envolvidas em uma determinada tarefa cognitiva, por exemplo. Em outro extremo, temos técnicas para registrar a atividade de um único neurônio. Microeletrodos podem ser inseridos próximos à célula nervosa para gravar o seu funcionamento. Observar o comportamento de neurônios isolados pode ser interessante por vários motivos. Por exemplo, cientistas notaram que alguns neurônios respondem de maneira extremamente seletiva a tipos específicos de estímulos (Iacoboni & Dapretto (2006)).

Entretanto, ambos os extremos são inadequados para fornecer informações a respeito de *como* as atividades cerebrais são executadas. As técnicas de mapeamento revelam padrões de conectividade e informações sobre a organização funcional do cérebro (quais áreas estão associadas à linguagem, memória, tomada de decisão etc.), mas possuem uma resolução baixa demais para entendermos como essas áreas cerebrais operam em função dos circuitos neurais que a compõem. As técnicas que registram as atividades de neurônios individuais, por outro lado, fornecem insights sobre funções importantes relacionadas a neurônios específicos, mas não nos mostram como o cérebro coordena grandes grupos de células para executar uma tarefa cognitiva complexa. Por exemplo, nós sabemos que a área de Brodmann está associada à capacidade de compreender e produzir linguagem, mas não sabemos como essa estrutura se articula internamente para processar informações, porque não temos tecnologia o suficiente para isso (e talvez nunca venhamos a ter).

As únicas ferramentas que temos para estudar as funções cerebrais na escala intermediária necessária são modelos computacionais. Os cientistas cognitivos chamam esses modelos de redes neurais artificiais, ou redes conexionistas. Esses modelos matemáticos abstraem da complexidade biológica envolvida em redes neurais reais, resguardando alguns aspectos importantes do funcionamento delas e dos neurônios que as compõem. O objetivo, como ocorre em qualquer modelo, é o de simular algumas operações tendo o controle de variáveis que seriam impossíveis de controlar numa situação real.

Uma rede neural artificial é composta de unidades de processamento, ou nós, que imitam algumas características de neurônios reais (ver Figura 4.1). Basicamente, um neurônio possui três partes, um corpo celular (por onde ele recebe sinais de outros neurônios), um axônio

(prolongamento por onde ele mesmo envia sinais elétricos) e dendritos (ramificações de saída que se conectam com células vizinhas). A função primordial de um neurônio é a de receber e enviar sinais elétricos. Então, é precisamente isso que o nó de uma rede conexionista faz. Ele recebe inputs elétricos de outros nós e, se esses inputs (I_j) somarem um dado valor específico, o neurônio dispara um sinal de saída.

Esse valor é chamado de limiar de ativação (L). Normalmente, ele é atribuído de maneira aleatória aos nós de uma rede, mas pode ser alterado conforme desejado. Outra característica simulada pelos nós é a de que neurônios podem enviar sinais excitatórios e inibitórios. Um sinal excitatório faz com que o neurônio receptor se aproxime de seu limiar de ativação, e um sinal inibitório faz o inverso. Essa característica é, numa rede neural artificial, representada pelos pesos (P_i) associados aos sinais. Um peso positivo corresponde a um sinal excitatório, e um peso negativo a um sinal inibitório. Se a somatória de todos eles (função de ativação) atingir o limiar de ativação do nó, ele dispara. O sinal disparado é idêntico para todos os nós que se conectam à sua saída.

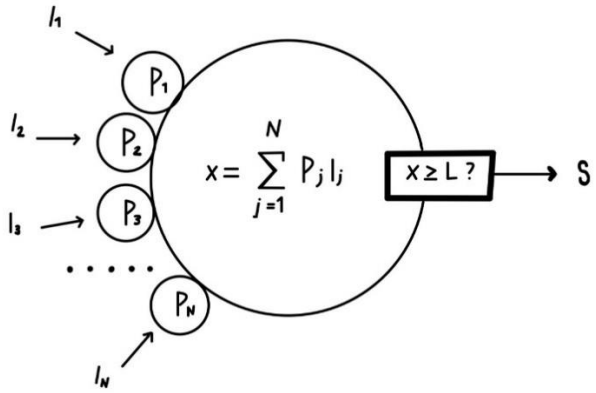

Figura 4.1 - Esquema de uma unidade de processamento de redes neurais artificiais (adaptado de Bermúdez (2014)).

Uma rede é composta de vários nós, formando fileiras ou camadas interconectadas. Os nós de uma mesma camada, porém, não se conectam. Há três tipos de camadas. Há a camada de entrada (que recebe input de fora da rede), a camada de saída (que emite o output da rede), e as camadas ocultas (que ficam entre a camada de entrada e a de saída). Não há limite para o número de camadas ocultas que uma rede pode ter, podendo ser, inclusive, 0. Redes que não possuem camadas ocultas, no entanto, são mais limitadas em termos do que elas podem fazer e aprender.

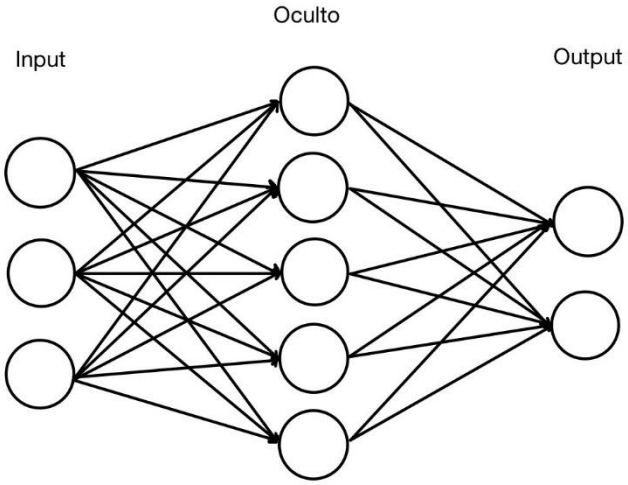

Figura 4.2 - Exemplo esquemático de uma rede neural artificial.

Uma das razões pelas quais os modelos de PDP se tornaram populares é a sua plausibilidade biológica. Ao contrário dos modelos clássicos – inspirados na arquitetura da máquina de Turing – os modelos de PDP são inspirados em neurônios. Todas as operações que ocorrem em redes neurais artificiais simulam operações que reconhecidamente ocorrem em redes de neurônios reais. A TCM, por exemplo, postula a existência de operações formais sobre estruturas simbólicas. Mas inúmeras questões práticas imediatamente se colocam diante da teoria, como "Em que nível fisiológico os símbolos são instanciados?" e "Quais mecanismos físico-químicos implementam operações simbólicas?". Esse tipo de questionamento não é considerado pelos defensores do PDP, pois a arquitetura cognitiva proposta pelos seus modelos é facilmente mapeada no cérebro.

Mas a plausibilidade biológica não é a única razão pela qual os modelos de PDP vêm angariando tantos adeptos. Redes relativamente simples são capazes de simular tarefas cognitivas relativamente complexas, a partir de poderosos algoritmos de aprendizagem. Sem entrar em muitos detalhes, veremos alguns exemplos de como isso ocorre.

4.1.2 ALGORITMOS DE APRENDIZAGEM EM REDES NEURAIS ARTIFICIAIS

Uma propriedade interessante de redes neurais artificiais é a de que elas são capazes de simular qualquer operação booleana. Basta ajustarmos corretamente os pesos dos sinais e os valores dos limiares de ativação dos nós. Uma porta OR, por exemplo, recebe dois valores de entrada e emite um valor de saída, de modo que a porta responde com FALSO quando as duas entradas são iguais a FALSO, e responde com VERDADEIRO em todos os outros casos. Uma única unidade de processamento é suficiente para implementar essa função. Assumimos que

todos os sinais, de entrada e de saída, variam no conjunto {0, 1}, e interpretamos 0 como FALSO e 1 como VERDADEIRO. Assumimos, também, que os pesos dos sinais de entrada são iguais a 1. Nesse caso, basta que o limiar de ativação do nó seja igual a 1. Quando ao menos um dos inputs for 1, o valor da função de ativação atinge o limiar, fazendo com que a unidade dispare 1. A unidade responde com 0 apenas na situação em que ambas as entradas são iguais a 0.

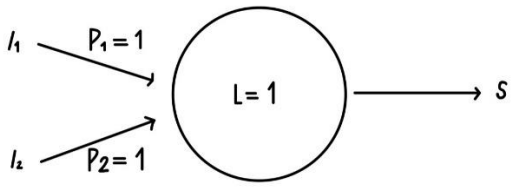

Figura 4.3 - Rede neural que implementa a porta lógica OR (adaptado de Bermúdez (2014)).

Mas, suponha que o output do nó seja diferente do output desejado, de modo que os pesos e/ou o limiar de ativação precisem ser alterados. Existe algum algoritmo que corrija esses valores? Rosenblatt (1958) descobriu um modo de solucionar esse problema, e chamou seu algoritmo de *perceptron convergence rule*. O que esse algoritmo faz é reduzir gradativamente o erro da rede, de modo que, se a regra for aplicada por um número suficiente de vezes, o output atual da rede será igual ao output desejado.

É importante observar que há dois tipos de aprendizado em redes neurais artificiais. Há o aprendizado *supervisionado* e o *não supervisionado*. No aprendizado supervisionado, a rede é "informada" sobre o grau de erro que ela cometeu em sua última implementação, de modo que os nós podem ser reconfigurados com base nessa informação. No caso do aprendizado não supervisionado, a rede não recebe nenhum tipo de feedback. Todos os métodos de aprendizado que veremos aqui serão métodos de aprendizado supervisionado.

A primeira coisa a se fazer é calcular o nível de *divergência* da saída da rede, i.e., em que grau o output atual diverge do output desejado. Chamemos esse número de δ:

$$\delta = \text{OUTPUT DESEJADO} - \text{OUTPUT REAL}$$

Se o output da rede é diferente do output desejado, δ é não nulo, e então poderemos mudar seu limiar de ativação ou reconfigurar os pesos atribuídos aos sinais de entrada. Essas duas opções são expressas pelas fórmulas:

$$\Delta L = - \alpha \times \delta$$

$$\Delta P = \alpha \times \delta \times I,$$

onde ΔL é a variação do limiar de ativação, ΔP é a variação do peso do sinal, α é um parâmetro que determina uma taxa de aprendizado da rede, e I é o valor do input. Suponha que δ tenha valor positivo. Significa que o output da rede foi menor do que o output desejado. Assim, a primeira fórmula fornecerá um valor negativo (assumindo que o parâmetro α tem valor positivo, o que geralmente é o caso), o que significa que o limiar de ativação deverá ser reduzido. Ao mesmo tempo, o peso do sinal de entrada deverá ser aumentado (ΔP positivo). O que acontecerá, segundo a *perceptron convergence rule*, é que o valor de δ se aproximará de 0, o que reflete o aprendizado da rede (que é nada mais do que o fornecimento do output adequado frente a um input específico).

Um caso concreto pode tornar o algoritmo mais claro. Suponha uma rede composta de um único nó, cuja tarefa é calcular a função NOT. Ela receberá apenas um input, cujo valor irá variar dentro do conjunto {0, 1}, e emitirá um output que varia também dentro do conjunto {0, 1}. Suponha que o peso do sinal de entrada seja P = – 0, 6, e o valor do limiar de ativação seja L = 0, 2.

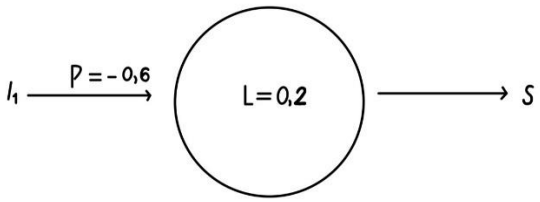

Figura 4.4 - Rede neural que implementa incorretamente a porta lógica NOT (adaptado de Bermúdez (2014)).

Quando I = 1, o valor da função de ativação será 1 × –0,6 = –0,6. Esse valor é abaixo do limiar e, portanto, o output é nulo (zero). Nesse caso a rede responde corretamente. Mas, quando o sinal de entrada é I = 0, o valor da função de ativação é 0 × –0,6 = 0. Como o limiar é 0,2, a rede também emite 0 como saída, onde era esperada emitir 1. Aplicando a regra de convergência, temos δ = 1. Suponha α = 0,5. Então ΔL = – 0, 5 × 1 = – 0,5. Temos ΔP = 0,5 × 1 × 0 = 0. As fórmulas nos dizem que o limiar de ativação deve ser reduzido em 0,5, enquanto que o peso do sinal de entrada é mantido inalterado.

Considere o novo limiar L = –0, 3. Quando o sinal de entrada é I = 1, o valor da função de ativação será 1 × –0,6 = –0,6, abaixo de L e, portanto, o output é, corretamente, 0. Quando a entrada é igual a 0, a função de ativação é calculada como 0 × –0,6 = 0, acima do limiar e, portanto, a rede dispara, corretamente, 1. Uma única aplicação do algoritmo foi suficiente para corrigir a rede nesse caso.

Entretanto, algumas funções booleanas não podem ser implementadas por redes tão simples. Considere a função XOR (ou *ou exclusivo*). Essa função binária emite VERDADEIRO somente nas situações em que apenas um dos disjuntos é verdadeiro, e emite FALSO em todas as outras situações (ou seja, quando ambos os disjuntos são verdadeiros e quando ambos são falsos). Se tentarmos construir uma rede que implemente a função XOR, e que seja composta de apenas uma unidade de processamento (como fizemos com as portas OR e NOT, por exemplo), ela terá que ser tal que, quando apenas o primeiro disjunto I_A emitir 1, ela emitirá 1. Ou seja, a rede precisa ser configurada de modo que $I_A \times P_A \geq L$. O mesmo vale para o caso em que apenas o disjunto I_B emitir 1. Desse modo, porém, será inevitável que quando $I_A = I_B = 1$, a saída seja 1, pois a soma de $I_A \times P_A$ e $I_B \times P_B$ é, necessariamente, maior que L.

Minsky e Papert (1969) provaram que redes compostas de apenas uma unidade de processamento são capazes de implementar somente funções booleanas linearmente separáveis:

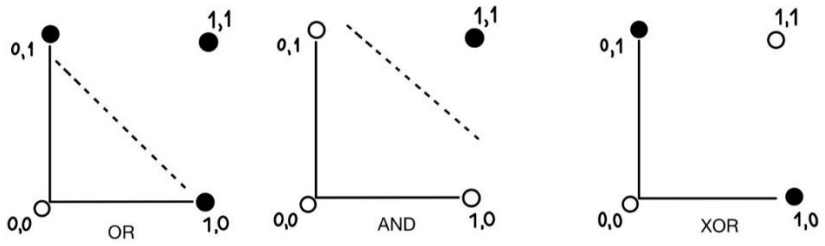

Figura 4.5 - Representação gráfica das operações lógicas OR, AND e XOR. Círculos hachurados representam saída igual a 1, círculos em branco representam pontos em que a saída é 0 (adaptado de Bermúdez (2014)).

Funções linearmente separáveis são aquelas cuja representação gráfica pode ser dividida em duas com uma única linha reta, que separe os pontos em que as entradas resultam 1 dos pontos em que as entradas resultam 0. Isso não pode ser feito com o gráfico que representa a porta XOR. No caso das funções binárias, temos, dentre 16, apenas 2 que não são linearmente separáveis. Mas, no caso das funções ternárias, por exemplo, temos 256 funções diferentes, dentre as quais apenas 104 são linearmente separáveis (Bermúdez (2014)). A diferença só aumenta com outras funções *n-árias* com $n > 3$. A imagem abaixo mostra uma rede capaz de implementar a função XOR.

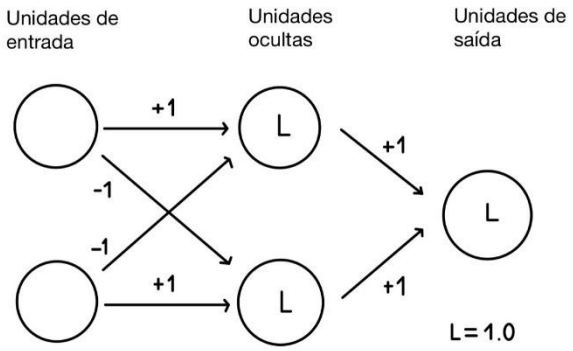

Figura 4.6 - Rede neural que implementa a função XOR (adaptado de Dawson (2003)).

Redes multicamadas podem implementar qualquer função booleana, sejam elas linearmente separáveis ou não. Na verdade, redes multicamadas podem implementar qualquer função computável (idem). O problema com essas redes é que não havia, até então, nenhum algoritmo de aprendizado disponível que fizesse uma rede sair de uma configuração aleatória de pesos e limiares em direção a uma configuração que fornecesse os outputs corretos frente aos inputs adequados. A *perceptron convergence rule* funciona com redes de apenas uma camada porque nós sabemos o output que elas devem fornecer. Mas, com redes multicamadas, não há como saber de antemão qual o output ideal para cada unidade que compõe as camadas ocultas. Sem essa informação, não conseguimos corrigir a configuração dessas unidades. Saber a função que a rede tem que executar é insuficiente, pois isso nos diz apenas o output que as unidades de saída precisam fornecer, mas não nos diz nada a respeito do que as unidades intermediárias precisam fazer. É um *problema de atribuição de crédito* (Dawson (2004)). O que nos falta é um método para atribuir o grau de responsabilidade que cada nó oculto possui com relação ao erro da saída.

Em 1986, Rumelhart, Hinton e William (1986) publicam um artigo com um método de aprendizado para redes multicamadas. Quando as unidades de entrada são excitadas com um input específico, a camada de saída fornece um *vetor de níveis de ativação*. Esse vetor pode ser subtraído do vetor de níveis de ativação desejado, gerando um novo vetor que nos informa o nível de erro de cada unidade de saída da rede. Se o nível de ativação de uma unidade de saída está abaixo do esperado, isso significa que houve pouca ativação nas camadas ocultas e, analogamente, se o nível de ativação de uma unidade de saída está acima do esperado, significa que houve um excesso de ativação nas camadas anteriores. Como vimos, isso ainda é insuficiente para corrigir os pesos das conexões entre unidades de camadas ocultas. No entanto, Rumelhart, Hinton e William encontraram um modo de estimar o grau de responsabilidade que as unidades

ocultas possuem com relação aos erros da camada de saída. Com base nesse grau de responsabilidade, os pesos das conexões das camadas ocultas são sucessivamente reconfigurados até que o erro da camada de saída seja reduzido a 0, ou a um valor próximo disso. Os detalhes matemáticos não serão expostos aqui, mas a ideia é a de que esse método de aprendizado toma o sentido inverso do sentido normal da ativação das unidades da rede. Enquanto que a ativação da rede se espalha sempre para frente, a correção dos erros se espalha para trás. A primeira camada a ser corrigida é a que se conecta com a camada de saída. Depois, a camada anterior a ela, e assim por diante. Por isso, o método ficou conhecido como *backpropagation error* (também conhecido como *regra delta generalizada*).

Em geral, muitos ciclos de treinamento são necessários para que uma rede atinja seu ponto ótimo de outputs corretos, também porque sua configuração de pesos iniciais é, frequentemente, aleatoriamente atribuída. Mas, quando isso acontece, a rede estaciona em uma configuração estável, capaz de fornecer outputs corretos para inputs que não apareceram em sessões de treino anteriores. Ou seja, a rede é capaz de fazer generalizações acertadas para estímulos novos. Utilizando esse método de aprendizado, pesquisadores obtiveram resultados impressionantes com modelos que simulam tarefas cognitivas das mais diversas. Gorman e Sejnowski (1988) criam um modelo de PDP que analisa e diferencia sinais de sonar provindos de dois alvos parecidos (um cilindro de metal e uma pedra de formato análogo) embaixo d'água, com precisão de até 90,4% em inputs não contidos nas sessões de treino (resultado comparável ao de especialistas humanos nessa tarefa). Hinton (1986) desenvolve um modelo capaz de analisar árvores genealógicas e aprender termos e relações de parentesco, completando proposições como "x é _____ de y e z". O modelo também responde corretamente para inputs não contidos em sessões de treino e, o que é mais interessante, captura informações que não estão explicitamente codificadas no input (Marcus (2001)), como a geração e ramificação familiar a qual uma pessoa pertence. Na Seção 4.1.3, veremos mais detalhadamente um modelo estudado por Rumelhart e McCleland (1986) que reproduz o padrão de aprendizado de crianças em fase de aquisição do passado simples do inglês. Todos esses modelos possuem grande impacto nas ciências cognitivas, não porque sejam sérios candidatos a modelar os processos cerebrais que executam essas funções, mas porque abrem uma possibilidade teórica alternativa ao computacionalismo clássico de entender os processos cognitivos.

4.1.3 REDES NEURAIS E A AQUISIÇÃO DA LINGUAGEM

Nossa conversa sobre os modelos de processamento distribuído paralelo começou com algumas considerações a respeito da abordagem de explicação de cima para baixo adotada pelo computacionalismo clássico e sobre como as redes neurais artificiais constituem uma alternativa a essa perspectiva teórica. Os defensores dos modelos de PDP argumentam a favor de uma abordagem científica menos hierárquica, na qual não há uma teoria mais privilegiada do que a outra, e segundo a qual todas coevoluem de maneira convergente e complementar. Desse modo, ao contrário do que muitos defensores do computacionalismo clássico argumentam, as categorias, princípios e generalizações destacados pela teoria representacional da mente poderiam estar equivocadas. Essa possibilidade se torna ainda mais concreta quando modelos de processamento de informação baseados em redes conexionistas se revelam capazes de implementar funções simulando uma arquitetura cognitiva completamente estranha ao simbolismo clássico.

Por exemplo, um princípio básico da psicologia de crenças e desejos é que, frequentemente, nosso comportamento é norteado por regras que são internamente representadas em nossas mentes. Assim, alguém que dirige na Inglaterra conduz o veículo segundo a norma de que o motorista fica sempre do lado esquerdo da rua. De acordo com o simbolismo clássico, a regra precisa ser, de algum modo, mentalmente expressa e causalmente eficaz, de modo que o comportamento do motorista seja sensível a ela. A TCM caracteriza essa regra como uma proposição da linguagem do pensamento que expressa algo como *O motorista deve dirigir sempre do lado esquerdo da rua*. Mesmo quando ele não esteja entretendo essa regra de maneira consciente, ela é explicitamente representada em seu cérebro. O ponto é que, de acordo com a TCM, o que explica que nosso comportamento C é conforme uma regra R é a própria existência de R em nossas mentes.

Essa ideia pode ser aplicada a diversos domínios cognitivos. A aquisição da linguagem, por exemplo, é um deles. Muitos autores tratam o aprendizado de uma língua como o domínio de regras a respeito de como arranjar e concatenar símbolos (Chomsky (2002)). Fodor (1975), por exemplo, argumenta que aprender o significado de um termo da língua natural nada mais é do que definir uma regra de verdade (*truth rule*) que estabelece uma relação de equivalência extensional entre o termo aprendido e um termo da linguagem do pensamento (ver Capítulo 1). Entender uma sentença do português requer uma tradução dos termos da sentença para os termos da linguagem do pensamento segundo as regras de verdade que aprendemos no curso de nosso desenvolvimento linguístico. Em linhas gerais, é isso que um computador faz para entender os comandos fornecidos por um humano em linguagem de programação de alto nível. Através de compiladores, comandos escritos em uma linguagem amigável para a compreensão humana

são traduzidos para a linguagem de máquina, composta apenas de 0's e 1's, para que a ordem possa ser executada.

Muitos fenômenos linguísticos são estudados a partir dessa perspectiva baseada em regras. O aprendizado do passado simples do inglês (*past tense*) é um fenômeno que, aparentemente, se encaixa bem nessa abordagem teórica. Os verbos do passado simples do inglês são divididos entre os regulares e os irregulares. A regra que comanda a transformação de um verbo regular do presente para o passado é, simplesmente, a adição do sufixo '-ed'. Por exemplo: 'walk – walked', 'represent – represented', 'develop – developed' etc. O que o falante precisa reconhecer é a raiz da palavra e a regra de adição do sufixo '-ed'. Os verbos irregulares, por outro lado, não possuem nenhuma regra clara a respeito de como devem ser transformados. Por exemplo: 'bring – brought', 'give – gave', 'sing – sang' etc. Devido a essas particularidades, a aquisição do passado simples do inglês em crianças nativas possui três estágios bem delineados (Bermúdez (2014)).

No primeiro estágio, apenas os verbos irregulares mais comuns aparecem em seu discurso. A principal hipótese é a de que esses verbos (e também os regulares) são aprendidos por imitação. Nessa etapa, as crianças tendem a não fazer generalizações a partir das palavras que foram adquiridas e, portanto, não cometem muitos erros.

Na segunda etapa, as crianças adquirem uma quantidade muito maior de verbos no passado simples (sendo que a grande maioria dos verbos é regular), aprendendo também a regra de transformação que consiste na adição do sufixo '-ed' à raiz do verbo. Nessa etapa, as crianças são capazes de fazer generalizações sobre palavras inventadas ('rick – ricked') e, além disso, cometem erros com verbos irregulares que haviam sido corretamente aprendidos ('give – gived'). Esses erros são chamados de erros de super regularização (*over-regularization errors* (Bermúdez (idem))).

Na última etapa, os erros de super regularização são reduzidos, e a performance linguística melhora tanto com os verbos irregulares quanto com os regulares.

Pinker e Prince (1988) ilustram, por meio de um modelo psicológico baseado em operações simbólicas, como as três etapas da aquisição do passado simples do inglês poderiam ser explicadas. De acordo com esse modelo, possuímos dois centros de processamento de informação cujo input é a raiz dos verbos do inglês. Um dos centros de processamento é a *memória associativa*, responsável por armazenar o passado simples de verbos irregulares. Esse módulo não aplica nenhuma regra que possa ser generalizada, apenas cria associações entre a raiz de um verbo e sua forma no passado simples. O outro centro, chamado de *módulo de hipóteses candidatas*, armazena e aplica uma regra generativa que adiciona o sufixo '-ed' à raiz do

verbo. O comportamento linguístico final do falante é função da interação competitiva entre esses dois módulos – cujos processos possuem pesos relativos construídos conforme o histórico linguístico experienciado pelo indivíduo.

Na primeira etapa do desenvolvimento linguístico, a criança desconhece as regularidades que definem as flexões verbais e, portanto, as formas dos verbos no passado que são utilizadas por ela são aplicadas apenas via memorização. Basicamente, nessa etapa, o output linguístico da criança é uma cópia do input linguístico que ela recebe e, assim, ela ainda não comete erros de super regularização porque esses erros não são testemunhados na experiência (e, além disso, ela ainda não formulou a regra que governa a flexão dos verbos regulares).

Em certo ponto, a regra do sufixo '-ed' é adquirida pelo módulo de hipóteses candidatas e, pela razão de que essa generalização é intensamente confirmada pela experiência (já que a maioria dos verbos do inglês é regular), ela é tomada como uma norma geral para produzir, a partir da raiz do verbo, seu passado simples. Há um período, portanto, em que verbos irregulares que haviam sido corretamente assimilados pela memória associativa são erroneamente transformados segundo a norma recém apreendida.

Na fase final do amadurecimento, o indivíduo percebe que tanto a forma irregular da raiz de um verbo quanto sua forma super regularizada correspondem a mesma coisa, viz., ao passado simples da raiz. Assim, por meio de um *princípio de singularidade* – o qual permite apenas uma única forma de flexionar a raiz de um verbo para cada categoria morfológica – as formas super regularizadas dos verbos são descartadas, pois precisam competir com formas irregulares que foram experienciadas ao longo do desenvolvimento linguístico (e, por isso, possuem maior peso relativo do que as formas super regularizadas, que não foram experienciadas em nenhum momento).

De acordo com Pinker (1984), a criança é capaz de representar mentalmente o input linguístico por meio de estruturas simbólicas, as quais denotam as características das expressões ouvidas em diferentes níveis de abstração (e.g., características lexicais, morfológicas, fonológicas etc.). Para todos esses níveis, generalizações podem ser formuladas, testadas e descartadas na fase de desenvolvimento. Um modelo psicolinguístico deve especificar quais generalizações são hipotetizadas pela criança e como elas interagem entre si para produzir o output linguístico.

O ponto aqui é o de como o computacionalismo clássico explica o comportamento (nesse caso, linguístico) recorrendo-se a regras mentais expressas por meio de uma linguagem simbólica inata, estruturada e extremamente complexa em termos sintáticos e semânticos. Mas, hoje em dia, os modelos de PDP abrem a possibilidade de entendermos nossa arquitetura cognitiva de maneira completamente avessa aos princípios teóricos que marcam a TCM. A seguir, veremos

de maneira resumida como Rumelhart e McClelland (1986) oferecem um modelo alternativo da aquisição do passado simples do inglês sem empregar nenhum tipo de regra ou expressões simbólicas.

Rumelhart e McClelland foram dos primeiros a trabalhar na área de modelagem cognitiva baseada em processamento distribuído paralelo. Os pesquisadores conseguiram treinar uma rede conexionista que recebe raízes de verbos do inglês como input e devolvem seu passado simples como output. A imagem abaixo representa a estrutura da rede.

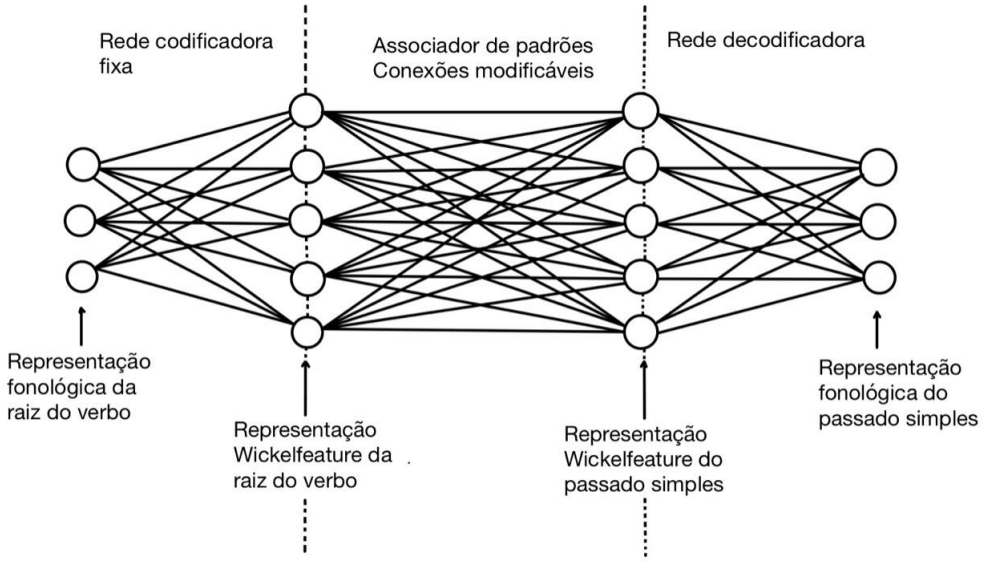

Figura 4.7 - Rede neural artificial projetada para aprender o passado simples do inglês (adaptado de Bermúdez (2014)).

Na verdade, essa rede é a união de três redes distintas. O aprendizado ocorre apenas na segunda rede, por meio do algoritmo *perceptron convergence rule*, visto na Seção 4.1.2. As outras duas redes são fixas, e fazem um trabalho de codificação do input da rede e decodificação do output da rede intermediária. A representação fonológica da raiz de um verbo é inserida como entrada da Rede Codificadora, e a representação fonológica do verbo no passado é devolvida como saída da Rede Decodificadora. Mas, o processamento da rede intermediária, que consiste, de fato, na flexão da raiz do verbo para o passado simples, ocorre sobre outros tipos representacionais, chamados *wickelfeatures*[55].

O que ocorre é que uma palavra é uma sequência ordenada de símbolos, e essa ordem se perde se não houver nenhum modo de codificar no input essa informação sequencial, pois os

[55] Esse nome se deve ao fato de que as ideias que Rumelhart e McClelland aplicaram ao modelo são inspiradas em conceitos propostos por Wickelgren.

nós da entrada representam apenas um conjunto não ordenado de fonemas (e.g., não seria possível distinguir a palavra 'ovo' da palavra 'voo', pois elas ativariam os mesmos nós na entrada da rede). Para superar esse problema, palavras são representadas como sequências de trigramas, chamados de *wickelphones*. Assim, por exemplo, a palavra do inglês 'strip' é codificada como $\{_{\#}S_t, {}_{s}t_r, {}_{t}r_i, {}_{r}i_p, {}_{i}p_{\#}\}$, onde o símbolo '#' é utilizado para demarcar as fronteiras inicial e final de uma palavra.

Entretanto, há pelo menos dois motivos para Rumelhart e McClelland não usarem wickelphones em seu modelo, um teórico e o outro prático. O motivo teórico, de acordo com Pinker (1984), é o que de um modelo projetado para fazer generalizações precisa das informações relevantes para ser capaz de formulá-las e, portanto, é importantíssimo que os primitivos representacionais do sistema sejam tais que essas informações possam ser extraídas e utilizadas. Ocorre que conjuntos de trigramas, nesse caso, não fornecem as informações detalhadas relevantes para detectar e formular os padrões pretendido pelos pesquisadores. A razão técnica é a de que o número de wickelphones possíveis seria da ordem de 43000 e, o número de conexões entre inputs e outputs das redes codificadora e decodificadora, da ordem de $(43000)^2$, um número que demanda recursos computacionais inviáveis.

Por isso, um wickelphone é diluído em um conjunto de *wickelfeatures*, que designa propriedades fonológicas da palavra. O segmento de letras 'ipt', por exemplo, corresponde aos wickelfeatures 'VogalDesvoziadaPlosiva' e 'AltaOclusivaOclusiva' (Pinker (1984)). Com alguns ajustes, o resultado final foi o de 460 unidades de processamento para cada camada das redes codificadora e decodificadora. A rede codificadora, portanto, transforma a representação fonológica da raiz do verbo em wickelfeatures, e a rede decodificadora transforma wickelfeatures em representações fonológicas. Como Steven Pinker observa, uma série de assunções linguísticas são feitas pelo modelo, as quais divergem significativamente das da teoria psicolinguística padrão.

A realização mais interessante do modelo é que não apenas a rede aprendeu a forma correta do passado simples de verbos regulares e irregulares do inglês, mas que a curva de aprendizado capturou as três etapas da aquisição mencionadas anteriormente.

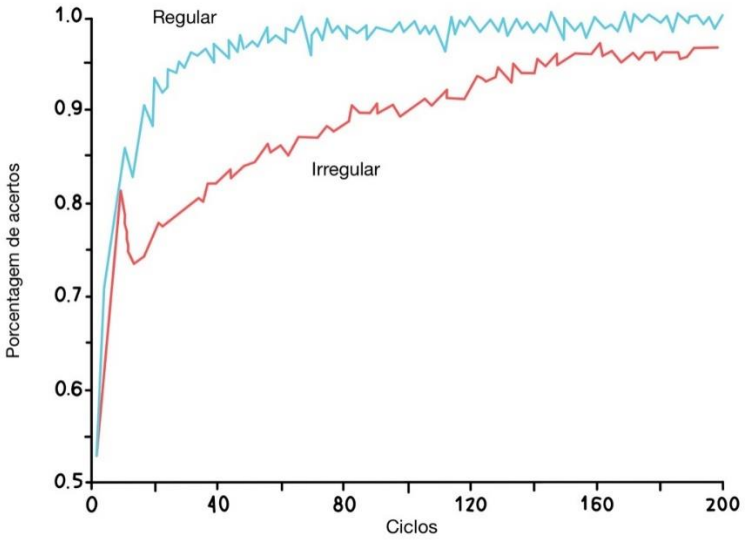

Figura 4.8 - Curva de aprendizado do passado simples do inglês da rede neural artificial projetada por Rumelhart e McClelland (adaptado de Bermúdez (2014)).

A rede aprende rapidamente verbos regulares e irregulares do inglês, mas, após o 11º ciclo de treinamento, há uma piora evidente na aplicação de verbos irregulares do inglês. Nesse momento, a rede artificial comete erros de super regularização, adicionando o sufixo '-ed' a verbos irregulares que haviam sido corretamente adquiridos anteriormente (e.g. 'give – gived'). Com o avanço do treinamento, porém, a rede retoma e supera sua performance anterior com verbos irregulares, ao mesmo tempo em que melhora também o domínio sobre verbos regulares.

A questão aqui é 'O conhecimento linguístico consiste de regras mentalmente representadas?'. Nosso comportamento linguístico é compatível com a hipótese de que há representações internas a respeito de regras sintáticas da língua, e que essas regras (i) são explicitamente representadas e (ii) são causalmente eficazes na produção linguística. Entretanto, o modelo se propõe a mostrar que o padrão comportamental detectado em um nível teórico pode não corresponder a um padrão isomórfico num nível de análise mais aprofundado. A rede conexionista de Rumelhart e McClelland reproduz o perfil de aprendizado exibido pelos falantes nativos do inglês, sem que qualquer regra do tipo 'adicionar sufixo '-ed' à raiz do verbo' seja codificada em suas unidades. O processo de aprendizado que conduziu a rede a um estado em que ela está disposta a responder corretamente a diferentes verbos do inglês nunca introduziu nenhuma regra de maneira explícita, apenas corrigiu os pesos entre as conexões das unidades da rede. O modelo abre a possibilidade de entendermos o conhecimento e o modo como as informações são representadas e processadas de uma maneira diferente do simbolismo clássico. Essa é a grande contribuição dos modelos de processamento distribuído paralelo.

4.1.4 CONSIDERAÇÕES A RESPEITO DA DIFERENÇA ENTRE MODELOS CLÁSSICOS E MODELOS DE PROCESSAMENTO DISTRIBUÍDO PARALELO

Como vimos anteriormente, os defensores dos modelos de PDP acreditam que uma mudança nos princípios da psicologia de senso comum pode emergir de baixo para cima. A ideia é que esses modelos têm o potencial de revelar de maneira concreta os mecanismos através dos quais o cérebro implementa nossas funções cognitivas, e que os insights fornecidos por eles podem nos forçar a rever e, em última instância, abandonar as categorias, mecanismos e explicações aos quais se agarram os teóricos que abordam o cérebro de uma perspectiva de cima para baixo.

Tendo analisado a rede criada por Rumelhart e McClelland, é interessante nos perguntarmos: a rede sabe transformar raízes de verbos do inglês para o passado simples? No que consiste esse conhecimento? O processo de aprendizado de uma rede neural consiste no ajuste dos pesos entre as conexões e dos limiares de ativação das unidades de processamento que a compõem. Quando esse ajuste ocorre com sucesso, a rede assume uma configuração estável, capaz de reconhecer padrões e realizar tarefas (nesse caso, fornecer o passado simples de um verbo do inglês). Mas, diferentemente do computacionalismo clássico, cujo conhecimento de uma regra (ou de um fato, ou de qualquer coisa) pode ser explicitamente representado através de um item discreto dentro do sistema (uma proposição da linguagem do pensamento), o conhecimento que a rede possui está *distribuído* nos pesos entre as conexões e nos limiares de ativação das unidades de processamento. Não há um parâmetro único na rede que corresponda a uma regra generativa que transforme raízes de verbos em seu passado simples.

O caráter distribuído do conhecimento da rede é inerente à sua própria estrutura. Todos os nós de uma camada estão conectados aos nós da camada subsequente, o que significa que todos os nós da camada anterior possuem a chance de influenciar as unidades da camada de saída. Essa característica de modelos de PDP possui um impacto profundo na forma como entendemos, por exemplo, as atitudes proposicionais. De acordo com a TCM, uma atitude proposicional corresponde a um item discreto dentro do sistema. A crença de que há uma regra para transformar verbos do inglês para o seu passado se reduz a uma proposição da linguagem do pensamento, a uma estrutura neural específica. Mas, isso não poderia ser dito de uma rede conexionista. Dizer que uma rede sabe transformar verbos é dizer algo sobre o complexo padrão de conectividade entre as suas unidades de processamento. Nada na rede corresponde a essa regra específica.

Alguns autores, portanto, argumentam que devemos caracterizar as atitudes proposicionais de maneira mais coerente com o caráter distribuído das redes de processamento distribuído paralelo. Uma possibilidade seria entender estados mentais como localizações num espaço multidimensional, onde cada localização equivale a matrizes de valores de ativação dos nós que compõem uma rede. Uma sequência de pensamentos seria o mesmo que uma trajetória nesse espaço multidimensional. Assim, as atitudes proposicionais seriam simplificações grosseiras a respeito dos estados neurofisiológicos que comumente caracterizamos como crenças, desejos, medos etc. Churchland escreve:

> Suppose that research into the structure and activity of the brain, both fine-grained and global, finally does yield a new kinematics and correlative dynamics for what is now thought of as cognitive activity. The theory is uniform for all terrestrial brains, not just human brains, and it makes suitable conceptual contact with both evolutionary biology and nonequilibrium thermodynamics. It ascribes to us, at any given time, a set or configuration of complex states, which are specified within the theory as figurative "solids" within a four- or five-dimensional phase-space. The laws of the theory govern the interaction, motion and transformation of these "solid" states within that space, and also their relations to whatever sensory and motor transducers the system possesses (CHURCHLAND, 1981, p. 129).

Nosso vocabulário seria uma simplificação dos processos distribuídos que de fato ocorrem dentro do crânio, não obstante seja conveniente do ponto de vista prático, uma vez que ainda carregam informação útil a respeito desses processos.

> According to the new theory, any declarative sentence to which a speaker would give confident assent is merely a one-dimensional *projection* – through the compound lens of Wernicke's and Broca's areas onto the idiosyncratic surface of the speaker's language – a one-dimensional projection of a four- or five-dimensional "solid" that is an element in his true kinematical state. Being projections of that inner reality, such sentences do carry significant information regarding it and are thus fit to function as elements in a communication system. On the other hand, being *sub*dimensional projections, they reflect but a narrow part of the reality projected. They are therefore unfit to represent the deeper reality in all its kinematically, dynamically and even normatively relevant respects (CHURCHLAND, 1981, p. 129).

Embora haja muitas críticas a respeito dos modelos de PDP que se propõem a explicar nossos processos mentais, a grande contribuição desses modelos é a de possibilitar a discussão sobre qual é a arquitetura mental implementada pelo cérebro. Provavelmente, ainda estamos longe de descobrir qual a opção correta, mas é inegável que ambas as abordagens requerem consideração por parte da ciência cognitiva contemporânea.

4.2 A COGNIÇÃO INCORPORADA E SITUADA

Um movimento recente nas ciências cognitivas tem ressoado a ideia de que a cognição é uma atividade realizada pelo corpo como um todo, e não apenas pelo cérebro. O cérebro, nesse caso, seria uma peça importante para entender nossas funções mentais, mas, ignorando os papéis do corpo, do contexto biológico, ambiental e cultural no qual o indivíduo se situa e age, não há esperança de uma compreensão satisfatória da cognição.

Esse movimento ficou conhecido como *cognição incorporada e situada* (CIS). No entanto, diferentemente do computacionalismo clássico, essa abordagem não desfruta de um consenso claro de asserções teóricas que a definam, de modo que seus defensores podem variar significativamente a respeito do que o termo *cognição* significa, e qual é exatamente o papel do corpo e do ambiente na explicação do comportamento (Shapiro & Spaulding (2021)). Alguns, por exemplo, rejeitam veementemente as noções teóricas empregadas pelo computacionalismo clássico, como *representação mental* e *algoritmo* (ver, por exemplo, Chemero (2009)). Outros aceitam essas noções, mas defendem uma reformulação a respeito de seu papel teórico nas ciências cognitivas (Barsalou (2016)). Em geral, os defensores da CIS desaprovam a ideia de que o cérebro seja um processador central, e que o comportamento seja o resultado da criação de modelos mentais internos e abstratos.

Nesta seção, veremos algumas das motivações que levaram os teóricos da CIS a mudarem de abordagem, e também alguns dos pilares teóricos em torno dos quais as discussões acerca dessa teoria ocorrem.

4.2.1 PROBLEMAS COM AS TEORIAS CLÁSSICAS DA COGNIÇÃO

O computacionalismo clássico forneceu insights teóricos importantes para o entendimento da cognição. É forçoso reconhecer, no entanto, que muitos aspectos importantes das atividades cognitivas, sobretudo cotidianas, realizadas por humanos e outros animais permanecem imunes a esse tipo de análise. Como Pfeifer e Bongard (2007) ressaltam, tarefas mundanas como a locomoção, a habilidade de reconhecer um rosto em uma multidão, de vestir uma roupa, de manipular uma chave de fenda ou folhear uma revista, que são flexível e elegantemente realizadas por nós, são dificilmente implementadas pelos modelos simbólicos de processamento.

Por exemplo, o modo preciso através do qual insetos se movem no ar em busca de alimento ou refúgio, descrevendo padrões complexos, é intrigante, levando em conta o tamanho diminuto de seus cérebros e os recursos computacionais limitados dos quais eles dispõem. Igualmente impressionante é o modo como um guepardo dispara em busca de uma presa, se adaptando aos obstáculos presentes no terreno acidentado. A locomoção, embora admirável dessa perspectiva, é realizada naturalmente e sem muito esforço por esses organismos. Entretanto, as primeiras tentativas da inteligência artificial de criar robôs que se moviam pelo cômodo e interagiam com objetos externos revelaram a dificuldade dessa tarefa. Shakey foi um dos precursores robóticos nesse sentido (Bermúdez (2014)). Suas capacidades consistiam em se mover pela sala através de rodinhas e executar ordens para mover caixas de lugar. Embora o ambiente fosse completamente controlado e as ordens fossem dadas em uma linguagem lógica extremamente simples em termos semânticos (composta apenas de conceitos como PORTA, CAIXA, PARAR, IR EM FRENTE etc.), Shakey demorava horas para processar os comandos que lhe eram passados, e apresentava pouquíssima flexibilidade em seus movimentos.

Uma das razões pela qual Shakey apresentava tamanha inabilidade motora é que o robô precisava criar modelos ou representações internas a respeito do ambiente no qual ele se encontrava, e esse modelo precisava ser constantemente atualizado conforme ele navegava pela sala e atuava sobre os objetos. Seu comportamento era fruto de um algoritmo sequencial que envolvia a "percepção" do ambiente por meio de sensores, a criação de "representações mentais" (que consistia na transdução dos elementos sensoriais para uma linguagem simbólica interna), o planejamento de uma ação por parte de um processador central, e a execução dessa ação através de uma ordem enviada ao sistema efetor. De acordo com a cognição incorporada e situada, há maneiras mais simples de conceber essas interações, que não requerem, necessariamente, a criação de modelos representacionais internos como mediadores.

A visão é um outro exemplo no qual modelos clássicos obtiveram algum avanço, mas está longe de se equiparar à capacidade adaptativa de olhos não artificiais. Formas artificiais de visão são bem sucedidas em ambientes controlados, em que as condições luminosas são constantes, as características geométricas do ambiente são escolhidas de maneira antecipada, a câmera está posicionada sempre no mesmo ponto e a fonte de energia é praticamente inesgotável. Quando essas condições são alteradas, o modelo falha (Pfeifer e Bongard (2007)).

O computacionalismo clássico teve sucesso precisamente nas tarefas em que humanos geralmente apresentam maior dificuldade, como a realização de cálculos complexos, a resolução de problemas abstratos e a avaliação posicional numa partida de xadrez. Outras formas de inteligência que requerem pouco esforço, mas maior flexibilidade, dificilmente são iluminadas

pela abordagem computacional. Pense nas estimativas que foram feitas desde a revolução tecnológica a partir de Alan Turing. Em 1950, Turing julgou que computadores evoluiriam a ponto de passar no jogo da imitação em meados dos anos 2000. Essa previsão não se concretizou. Na década de 60, um dos grandes objetivos da inteligência artificial era o de criar um tradutor que recebesse um texto em uma língua e entregasse esse mesmo texto em outra língua. Novamente, essa tarefa se mostrou extremamente árdua, e os melhores resultados estão muito distantes de exibir algo próximo da capacidade humana de lidar com a linguagem natural (Baker (2001)). Essas frustrações levaram teóricos a buscar alternativas ao computacionalismo clássico.

4.2.2 PERSPECTIVA EVOLUTIVA E MUDANÇA DE PARADIGMA INVESTIGATIVO NA COGNIÇÃO INCORPORADA E SITUADA

Uma das reflexões feitas pelos teóricos da CIS é que a cognição é fruto de um processo evolutivo e, portanto, suas propriedades foram selecionadas por apresentarem vantagens adaptativas. Além disso, como ocorre com qualquer característica fenotípica, o valor da cognição é condicionado ao meio que a selecionou. Considere os olhos. Em um mundo que não há luz, não há organismos com órgãos sensíveis a ela, porque esse órgão não ofereceria nenhuma vantagem adaptativa. Desse modo, é proveitoso investigar a cognição – seja a humana, seja a de outro animal – tendo em vista (i) o papel adaptativo que ela teve na vida do organismo e (ii) o ambiente no qual ela se destacou.

A respeito de (i), apenas muito recentemente usamos o cérebro e o corpo para fazer cálculos complexos e abstratos ou para encadear logicamente uma sequência longa de proposições. Primordialmente, a cognição foi útil aos nossos antepassados para que eles fossem capazes de encontrar comida, se locomover de maneira eficiente, identificar sinais de perigo e se comunicar com outros indivíduos do grupo. Se considerarmos formas de vida mais rudimentares, a cognição, para eles, possui um papel ainda mais primitivo.

Sobre o ponto (ii), é importante observar que o meio no qual um organismo se desenvolveu selecionou ativamente suas propriedades morfológicas e, por essa razão, influenciou diretamente também os aspectos de sua natureza cognitiva. Isso porque a cognição está, em muitos casos, se beneficiando e, ao mesmo tempo, *restrita* a essas características morfológicas. Por exemplo, Nagel (1974), em outro contexto, argumenta a favor da tese de que é psicologicamente impossível para nós, humanos, imaginar como é ser um morcego (ver Seção 4.2.4).

Duas considerações podem ser feitas a partir dessa breve reflexão. A primeira é que a cognição não pode ser compreendida sem levar em conta o papel do corpo. A cognição é parte do corpo, e foi moldada evolutivamente em harmonia com ele – i.e., não de maneira isolada. Esse ponto é importante porque, até o computacionalismo clássico, as ciências cognitivas tiveram uma tendência de tratá-la exclusivamente como um processo local, em que o *locus* físico é o cérebro. A cognição incorporada e situada, porém, favorece uma visão mais global, de acordo com a qual processos cognitivos devem ser entendidos como processos essencialmente corporais. Afinal, nossas capacidades cognitivas são adaptadas para o tipo de corpo que temos, para que possamos agir com ele de maneira eficiente no meio em que nos encontramos.

A outra consideração é metodológica. Ao invés de investigar processos cognitivos sofisticados, a CIS aborda de maneira mais incisiva processos mais simples, como interações sensório-motoras do organismo com o ambiente. Insetos, por exemplo, são uma grande fonte de inspiração para a cognição incorporada e situada. Eles são capazes de executar tarefas complexas a partir de mecanismos relativamente simples. A capacidade de entender e recriar em laboratório esses comportamentos, se possível, representa um grande avanço na compreensão da cognição. A cognição incorporada e situada, desse modo, se aproveita dos insights provenientes da biorobótica, na qual engenheiros e biólogos tentam imitar o comportamento desses organismos em laboratório. A ideia é que, partindo dos processos mais simples de cognição, um dia conseguiremos entender os processos mais sofisticados, como a linguagem e o raciocínio lógico. Conforme prevê o título de um artigo de um dos precursores da CIS, Rodney Brooks, '*Today the earwig, tomorrow man*?'.

4.2.3 BIOROBÓTICA E CIS

Um exemplo notável do papel do corpo em mecanismos cognitivos vem da forma como grilos fêmeas localizam grilos machos com base no som que eles emitem. Barbara Webb (1996) conduziu uma pesquisa na tentativa de replicar esse fenômeno – conhecido pelos biólogos como *fonotaxia*. Então, como as fêmeas encontram os machos através do som? À primeira vista, alguém poderia imaginar que elas precisariam de um considerável tratamento computacional do sinal sonoro que elas captam, distinguindo o som dos machos dos ruídos do ambiente, por exemplo, ao mesmo tempo em que situaria a origem desse som em algum tipo de mapa mental, após o que seu cérebro enviaria ordens motoras aos seus membros inferiores para saltar na direção do som. Entretanto, a solução desse problema consiste em algo bem mais direto. A resposta se encontra na *morfologia* do grilo.

Diferentemente de nós, os ouvidos do grilo estão localizados em suas pernas, conectados um com o outro por tubos traqueais, de modo que o som que chega em um ouvido pode chegar por meio de rotas diferentes – diretamente pelo ambiente, ou pelo tubo traqueal (tendo atingido o outro ouvido primeiro). A fonotaxia opera fundamentalmente por meio de duas propriedades morfológicas do grilo: tubos traqueais que estão mais próximos da origem do som vibram mais vigorosamente (o que informa a direção do grilo macho) e a vibração desses tubos controla diretamente os saltos do grilo fêmea.

Observe que não há nenhum tipo de mecanismo computacional de identificação do som emitido pelos grilos machos. Não há a criação de um mapa mental com a localização da origem do som, e também não há um processador central que gere comandos motores para guiar os saltos do grilo fêmea. A engenhosidade do processo está incorporada nas características morfológicas de seu corpo – um exemplo de *computação morfológica*. A ideia é que animais podem usar suas características biológicas para executar tarefas inteligentes, poupando recursos computacionais que, de outro modo, eles não teriam. Há uma interface direta entre *percepção* e *ação*. Ao contrário de Shakey, por exemplo, o comportamento do grilo fêmea não é intermediado por cálculos computacionais e modelos representacionais complexos. Na literatura, também encontramos o termo *enativismo* para designar casos como esse (Huto & Myin (2013), Pantaleão (2021)). Em geral, enativistas argumentam que a cognição emerge de, ou é constituída por, atividades sensório-motoras (Shapiro & Spaulding (2021)). A ideia, novamente, é eximir as ciências cognitivas da premissa de que o comportamento é sempre intermediado por representações internas.

Outro exemplo enativista é o robô Allen, construído por Brooks (1986). Allen possui uma arquitetura que, assim como o mecanismo de fonotaxia dos grilos, conecta inputs sensoriais diretamente com outputs comportamentais, sem a necessidade de haver um complexo processamento de informação intermediário.

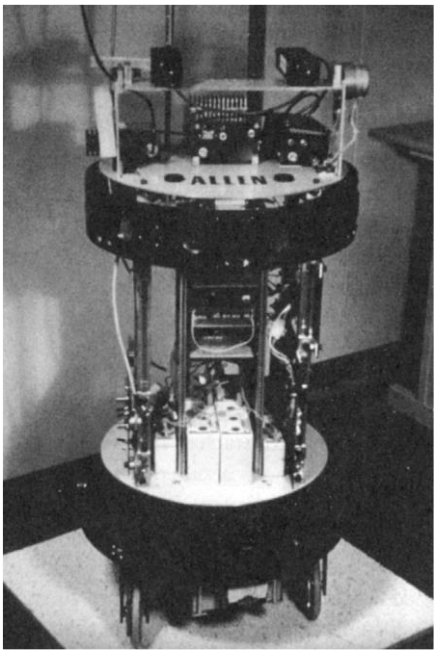

Imagem 4.1 - Robô Allen (Brooks (1986)).

Isso significa que o robô consegue interagir e navegar pelo ambiente de maneira ágil e flexível, ao contrário, por exemplo, de seu antecessor, Shakey. A chave é embutir no sistema mecanismos que operam de maneira responsiva, realizando ações simples, mas versáteis. Além disso, é possível construir subsistemas mais complexos, que integram esses mecanismos mais simples, compondo uma hierarquia de níveis. Esse tipo de organização é chamado de *arquitetura de subsunção*. O que é característico desse tipo de projeto é que cada subnível é construído de maneira independente, depurado e, depois, acoplado com o nível acima, de maneira acumulativa.

Alguns dos subsistemas de Allen são o *Avoid Objects* (responsável por evitar obstáculos do ambiente); o *Wander* (faz o robô vagar pela sala); e o *Explorer* (auxilia o robô a explorar o cômodo em busca de algum objetivo pré-estabelecido). O nível mais básico é o Avoid Objects, e o mais alto é o Explorer. Todos operam de maneira mais ou menos autônoma, embora certos níveis se sobreponham a outros em situações específicas. Por exemplo, se Allen estiver explorando o cômodo e precisar desviar de um obstáculo, o Avoid Objects irá operar de maneira forçada, mesmo estando abaixo na hierarquia de níveis. Além disso, de acordo com Brooks, todo subnível é um *sistema de produção de atividade*, o que significa que todos conectam diretamente *percepção* e *ação*. Todos operam cem por cento do tempo, de maneira paralela e sem a necessidade de um processador central. Essas características favorecem um comportamento mais ágil e flexível por parte do robô.

Entretanto, a maior diferença entre a arquitetura de subsunção e as IAs clássicas reside no modo como Allen *representa* o mundo no qual ele se situa e com o qual ele interage. Brooks afirma que não há uma representação central que reúna as informações coletadas pelos sensores,

formando um modelo unificado do mundo. As informações sensoriais captadas pelos subsistemas de Allen são muito fragmentadas e sequer estão todas conectadas em um circuito único. Não é possível identificar, no hardware do robô, um local que corresponda à "saída" da percepção.

O fato de que Allen não opera sobre modelos representacionais internos chama a atenção porque, embora esse seja um princípio de montagem que confere agilidade e versatilidade ao robô, há um grande debate em torno da questão de se esse tipo de mecanismo constitui inteligência genuína (Shapiro & Spaulding (2021)). Representações ocupam um papel teórico privilegiado nas ciências cognitivas, mesmo quando não há consenso sobre o que são essas entidades.

Na Seção 4.2.4, veremos como a CIS pode aceitar uma noção mais robusta de representação mental, ao mesmo tempo em que destaca o papel do corpo na articulação desse conceito teórico.

4.2.4 CONCEITUALIZAÇÃO

A conceitualização é a tese de que as características morfológicas de um organismo limitam quais conceitos ele pode adquirir ao longo da vida, de modo que características morfológicas distintas implicam maneiras distintas de entender o mundo (Shapiro & Spaulding (2021)). A ideia é que os blocos de construção de nossos conceitos ou categorias mentais são as experiências provenientes das diferentes modalidades sensoriais, como tato, olfato, visão etc. Se esse for o caso, organismos com modalidades sensoriais diferentes – qualitativa ou quantitativamente – possuem ferramentas cognitivas distintas para entender o mundo.

Novamente, a motivação aqui é a de incluir o corpo, o ambiente e o contexto social na explicação de nossas capacidades cognitivas. Classicamente, a cognição é considerada a etapa intermediária entre a percepção e a ação – o que alguns chamam de modelo sanduíche da cognição (Barsalou (2016)). Essa abordagem favorece uma visão modular de nossos processos cognitivos, considerando-os independentes (no sentido de que operam autonomamente) dos processos anteriores e posteriores a eles.

Por exemplo, Fodor (1975) argumenta a favor da tese de que o output de nossos processos perceptuais é convertido para a 'linguagem de máquina' de nosso cérebro, a linguagem do pensamento. De acordo com o autor, haveria um processo de *transdução* dos inputs específicos a uma modalidade particular (e.g. visão) para uma linguagem 'neutra', que serve de veículo para o processamento de informação no cérebro. A percepção visual de que chove é, ao

final, transcrita para uma proposição da linguagem do pensamento cujo conteúdo é algo como 'Chove'. Uma das vantagens teóricas dessa tese é que ela consegue explicar como informações provenientes de diferentes modalidades (e.g. visual, auditiva, linguística etc.) podem ser comparadas, cruzadas, confirmadas e confrontadas – porque, uma vez que elas são traduzidas para uma linguagem comum, elas se tornam *mensuráveis*. Confirmação, por exemplo, parece ser uma relação entre *proposições*. Como o som da chuva pode confirmar a hipótese de que está chovendo? De acordo com a teoria, essa relação de confirmação ocorre porque a percepção do som envolve a criação de uma estrutura proposicional na mente que torna mais provável a verdade de outra proposição mental cujo conteúdo é de que chove.

O ponto é que, de acordo com o computacionalismo clássico, informações essencialmente *modais* (dos sentidos) são traduzidas para um veículo essencialmente *amodal*. Essa maneira de ver a cognição torna conveniente a abordagem de estudar processos cognitivos separadamente dos inputs sensoriais e outputs comportamentais. Defensores da CIS, por outro lado, argumentam que a cognição não será satisfatoriamente compreendida se esses outros aspectos não forem levados em consideração (Broens & Gonzalez (2006)). De fato, muitas evidências empíricas apontam para a ideia de que nossos conceitos são, na verdade, modais (Barsalou et al (2003)). Antes de vermos as evidências, porém, discutiremos um pouco o fundamento teórico que permite uma interpretação dos resultados empíricos.

4.2.5 MODELOS DE PDP E A COGNIÇÃO INCORPORADA E SITUADA

Muitos teóricos da cognição incorporada e situada utilizam modelos de processamento distribuído paralelo para entender como o cérebro representa nossas categorias mentais. Vimos na seção anterior que redes neurais artificiais são capazes de realizar tarefas complexas quando ajustadas corretamente. Essas tarefas incluem, por exemplo, o reconhecimento de padrões. E o que é a aplicação de um conceito se não o reconhecimento de um certo padrão? Quando identificamos um objeto como um pássaro, por exemplo, o que estamos fazendo é verificando se as características daquele animal correspondem a uma lista de características típicas que atribuímos aos pássaros – tem bico, asas, penas, voa etc. Quanto mais características do animal constarem nessa lista, maior a chance que teremos de classificá-lo como um pássaro (Murphy (2001)). Redes neurais são um excelente modelo nesse sentido, porque os nós da rede podem corresponder aos itens dessa lista, e quanto mais nós forem ativados (ou seja, quanto mais características típicas forem observadas), maior a chance da rede classificar o input na categoria relevante.

A ideia é que um conceito é um *atrator* da rede – uma configuração estável, porém, dinâmica, que ela assume quando identifica um padrão (Barsalou (1999)). Por exemplo, quando vemos um pintassilgo, nosso conceito dinâmico de pássaro é ativado. Quando vemos uma galinha, nosso conceito de pássaro, ou seja, o mesmo atrator da rede, também é ativado, porém com diferenças que são sensíveis ao contexto (um pintassilgo é um pássaro muito mais típico do que uma galinha, que não voa, por exemplo). Diferentes contextos podem distorcer a ativação da rede, o que explica por que nossos conceitos evidenciam certas características em detrimento de outras de acordo com a situação específica em que são aplicados (e.g. a característica 'pingar' é ativada quando pensamos em uma bola de basquete, mas, se estivermos naufragados no mar e vermos uma bola de basquete, a característica 'flutuar' provavelmente será mais fortemente ativada).

Diferentemente da abordagem computacionalista, que considera nossos conceitos estruturas discretas, os defensores da CIS veem nossos conceitos como configurações dinâmicas em uma rede de neurônios. Além disso, nossos conceitos não seriam amodais, ou neutros, com relação às categorias sensoriais. Pelo contrário, o uso de um conceito pode ser visto como uma simulação sensório-motora da experiência que tipicamente temos quando nos encontramos com os objetos dessa categoria.

Isso significa que um conceito geralmente não é concebido contra um pano de fundo branco, de maneira isolada. Conceitos geralmente são simulados levando em conta os contextos típicos nos quais eles se aplicam. Quando imaginamos uma bicicleta, imaginamos como é andar nela, como nossos sentidos são afetados nesse contexto específico etc., o que nos permite fazer inferências úteis sobre objetos, pessoas e situações que podem ser encontradas quando andamos de bicicleta.

4.2.6 EVIDÊNCIAS EMPÍRICAS ACERCA DA MODALIDADE DE NOSSOS CONCEITOS

Um exemplo interessante a respeito da hipótese de simulação mental é encontrado em Shepard & Metzler (1971). No experimento, sujeitos observam duplas de figuras e precisam responder se elas correspondem ao mesmo objeto.

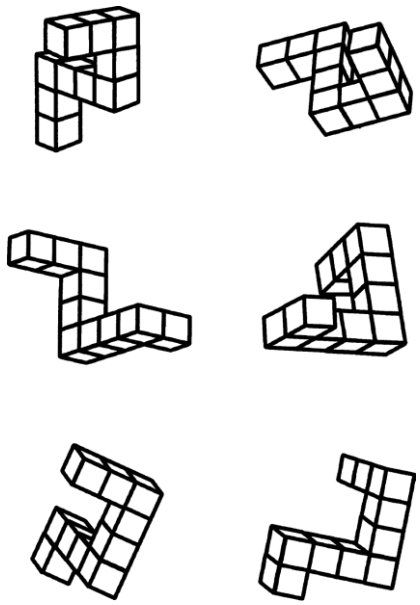

Figura 4.9 - Figuras utilizadas no experimento de rotação proposto por Shepard e Metzler (1971).

Nenhuma ordem explícita foi dada a respeito de como eles deveriam encontrar a resposta. Simplesmente, deveriam puxar uma alavanca se as figuras correspondessem ao mesmo objeto, e puxar outra alavanca se não correspondessem. O que é interessante é que o tempo de resposta dos participantes é diretamente proporcional à diferença no ângulo de rotação entre o par de figuras. Ou seja, quanto maior for a diferença entre os ângulos de rotação, mais demorada será a resposta. A principal hipótese é a de que, para realizar essa tarefa, participantes *simulavam mentalmente* como é rotacionar esse objeto, até que ele chegue na posição desejada. Essa simulação seria essencialmente modal, no sentido de que os circuitos sensório-motores utilizados no manejo de objetos como esses seriam os mesmos empregados na simulação, recriando a experiência que teríamos caso víssemos e tocássemos esses objetos de fato. Modelos clássicos de inteligência artificial traduziriam essas imagens para uma linguagem amodal, e fariam cálculos para testar a correspondência entre elas. Entretanto, o tempo de demora que custaria a execução desses cálculos não é uma função do ângulo de rotação entre as figuras (Bermúdez (2014)). O computacionalismo clássico, portanto, teria dificuldades em explicar os resultados desses experimentos.

Outra série de experimentos demonstra como nossos estados corporais são ativados quando processamos informação conceitual (Barsalou et al (2003)). Por exemplo, Tucker e Ellis (1998) demonstraram como a orientação de um objeto (e.g. esquerda e direita) interfere na resposta corporal dos participantes em função da (falta de) congruência entre os dois. Participantes respondiam se um objeto, e.g. uma panela, estava virada de cabeça para baixo ou não. A panela podia estar orientada para a direita (e.g. cabo voltado para a direita) ou para a

esquerda. A resposta dos participantes era dada através de dois botões, um botão esquerdo e um botão direito. O tempo de resposta era menor quando o botão a ser pressionado era congruente com a orientação do objeto.

Outro exemplo parecido (Glenberg e Kaschak (2002)) mostrou que a compreensão de uma sentença (e.g. 'abrir a secadora') também envolve ativação de estados corporais. Por exemplo, a ação de abrir uma secadora requer um movimento corporal específico, e o tempo de resposta dos participantes é menor se a resposta envolver um movimento congruente àquele ativado pela sentença. Assim, participantes levam mais tempo para responder a sentença 'fechar a secadora' se essa resposta envolver o movimento de abrir ou puxar algo.

A hipótese dos defensores da CIS é que esses experimentos mostram que o uso de conceitos ativa regiões específicas do cérebro, regiões que são geralmente ativadas quando experienciamos de maneira concreta objetos dessas categorias. Essas regiões são específicas a modalidades sensoriais e motoras. A conclusão é que o uso de um conceito é uma simulação sensório-motora que envolve o cérebro e o corpo, e que remete às situações que experienciamos no passado. Uma vez que essas experiências emergem da interação complexa entre o corpo, o ambiente e o contexto social, não é difícil entender como organismos diferentes compreendem o mundo de maneiras distintas.

4.2.7 COGNIÇÃO ESTENDIDA

Por fim, é pertinente mencionar que alguns autores defendem não apenas que a cognição é um processo corporal, que se estende além do sistema nervoso, mas também um processo que abrange objetos do ambiente, como calculadoras e celulares. Frequentemente delegamos tarefas cognitivas pesadas a esses objetos. Calculadoras fazem contas por nós, e celulares salvam números de telefone para economizarmos espaço em nossa memória interna. Em um sentido importante, eles são extensões de nosso sistema cognitivo.

Quando um jogador de baseball usa o taco para bater na bola, por exemplo, o taco funciona como se fosse uma extensão de seus membros. Talvez, (jogador + taco) possa ser visto como um sistema único, para fins didáticos ou psicológicos. No entanto, dificilmente alguém afirmaria que, durante o jogo, o taco literalmente faz parte do sistema motor do jogador. Mas é precisamente essa a ideia por trás da cognição estendida. No caso de aparelhos que auxiliam a cognição, a tese é que eles constituem parcialmente o sistema cognitivo composto de (indivíduo + aparelho (e.g. calculadora)).

Os defensores da cognição estendida argumentam que a delimitação espacial que fazemos do sistema cognitivo é simplesmente arbitrária. Não há, de acordo com eles, boas razões para restringirmos a cognição ao nosso crânio, nem mesmo ao nosso corpo. Considere o seguinte cenário. João sofre um acidente de carro. A batida é tão forte que seu sistema auditivo é danificado e ele para de escutar. Os médicos e engenheiros constroem um aparelho que irá substituir a parte danificada de seu sistema auditivo. Literalmente, os médicos irão remover essa parte e inserir a peça recém criada. Por que não dizer que essa peça, agora, faz parte do sistema auditivo de João? E se admitirmos que esse é o caso, pareceria arbitrário afirmar que é uma condição necessária que a peça esteja dentro de seu crânio. Suponha que a peça seja muito grande para ser embutida em seu ouvido interno, e precise ser conectada ao seu sistema nervoso através de micro transdutores de sinal, de modo que ela fique acoplada a João de maneira externa. Nesse caso, a peça não faria parte de seu sistema auditivo? Parece arbitrário, diriam os defensores da cognição estendida, negar que esse seja o caso.

Vale observar, porém, que a tese da cognição estendida não constitui uma rival da teoria computacional da mente. São teses consistentes. Alguém poderia defender que nosso cérebro executa computações sobre estruturas simbólicas *à la* Máquinas de Turing, ao mesmo tempo em que admite a ideia de que nossos computadores e celulares, junto com nosso sistema nervoso, constituem um grande sistema cognitivo, por onde a informação é representada, processada e armazenada.

4.2.8 DIFERENÇAS ENTRE A CIS E O COMPUTACIONALISMO CLÁSSICO

Como vimos, a cognição incorporada e situada é um movimento heterogêneo, que nem sempre concorda sobre as mesmas teses. A ideia chave dessa corrente teórica, porém, é que o corpo possui um papel muito mais proeminente na cognição do que julgam os teóricos do computacionalismo clássico. Vimos que os modelos computacionais clássicos têm dificuldade em explicar as funções simples que nós e outros organismos costumamos realizar no dia a dia. Ao contrário, os modelos clássicos se destacam precisamente nas tarefas em que humanos possuem dificuldade de realizar, como jogar xadrez e realizar cálculos longos e abstratos (§5.2.1). A partir da abordagem da cognição incorporada e situada, é possível entender como comportamentos relativamente complexos, como a locomoção e interação com objetos do ambiente, são realizados a partir de processos essencialmente corporais, sem a necessidade de um modelo representacional interno ou complexos processamentos de informação (§5.2.3).

Além disso, apresentamos algumas questões sobre as quais não há consenso dentro da CIS, como a questão da necessidade de haver representações internas ou não (§5.2.4). Os enativistas tendem a diminuir o papel das representações mentais nas explicações dos comportamentos dos organismos. Ao mesmo tempo, outros autores aceitam esses construtos teóricos, argumentando, no entanto, que representações possuem uma natureza modal, i.e., cujas características estão associadas às nossas modalidades sensoriais – em oposição à natureza amodal dos símbolos postulados pelo computacionalismo clássico. Vimos algumas evidências empíricas a favor da modalidade de nossas representações internas e como essa ideia está embasada nos modelos de processamento paralelo vistos na seção anterior.

Por último, apresentamos a hipótese da cognição estendida, de acordo com a qual processos cognitivos devem ser entendidos de maneira mais abrangente, envolvendo não apenas o corpo, mas os próprios objetos que auxiliam nossos processos inteligentes, como calculadoras e notebooks. Argumentamos que a hipótese da cognição estendida não constitui, necessariamente, uma adversária ao computacionalismo clássico, de modo que as duas hipóteses são consistentes entre si.

CONSIDERAÇÕES FINAIS

> [...] some of the most striking things that people do – 'creative' things like writing poems, discovering laws, or, generically, having good ideas – don't feel like species of rule-governed processes (FODOR 1975, p. 201).

A teoria computacional da mente constitui uma hipótese a respeito do mecanismo cognitivo subjacente ao que chamamos de *raciocínio*, ou, de maneira mais geral, processos racionais. De acordo com a TCM, o cérebro instancia estruturas simbólicas com conteúdo semântico e propriedades sintáticas, e manipula essas estruturas segundo regras computacionais. A ideia fundamental é que (i) os processos de manipulação e transformação de símbolos são sensíveis unicamente à sintaxe dessas estruturas, i.e., às suas propriedades formais, e (ii) as propriedades sintáticas dos símbolos mapeiam as relações semânticas entre eles, de maneira análoga ao que ocorre com os computadores modernos. Como vimos no Capítulo 1, as propriedades formais dos símbolos são instanciadas pelas propriedades físicas do veículo representacional – no caso do computador, por polaridades magnéticas, no caso do cérebro, por propriedades neurofisiológicas. O raciocínio é um tipo especial de processo físico – por que esses processos são chamados de racionais? Precisamente porque a sucessão causal de estados psicológicos se constrói em torno da relação semântica que existe entre eles.

É importante reconhecer, no entanto, que nem toda sucessão causal de eventos psicológicos preserva suas relações semânticas. Em outras palavras, nem todo evento mental é fruto de um processo meramente sistemático e racional. De fato, é possível e até provável que a grande maioria não o seja. Kahneman (2011), em sua obra *Rápido e Devagar*, apresenta vários experimentos psicológicos projetados para expor o quão irracionais são os seres humanos. Por exemplo, em um dos experimentos, Daniel Kahneman e Amos Tversky – seu amigo e colega de profissão – adulteraram uma roda da fortuna que possuía números de 1 a 100 para parar apenas nos números 10 ou 65. Kahneman ficava em frente a roda, girava-a e pedia a um pequeno grupo de estudantes que anotassem o resultado, que era sempre 10 ou 65. Depois, os participantes do experimento precisavam responder a duas perguntas:

– A porcentagem de nações africanas entre membros da ONU é maior ou menor do que o número que você acabou de escrever?

– Qual é a sua melhor estimativa sobre a porcentagem de nações africanas na ONU?

Conforme Kahneman reporta em seu livro:

> O giro de uma roda da fortuna – mesmo de uma que não esteja adulterada – não tem como fornecer informação útil sobre o que quer que seja, e os participantes de nosso experimento deveriam simplesmente tê-la ignorado. Mas não o fizeram. As estimativas médias dos que viram 10 e 65 foram 25% e 45% respectivamente (KAHNEMAN, 2011, p. 152).

Esse efeito psicológico é chamado de *efeito de ancoragem*. Se formos requisitados a fazer uma estimativa e esse pedido vier acompanhado de um valor, levaremos em conta esse valor mesmo que ele não forneça nenhuma informação útil para o problema, de modo que nossa estimativa tenderá a cair próxima dele. Daí a ideia de uma âncora. O experimento revela um caso em que há uma conexão causal entre estados psicológicos – o valor fornecido pela roda da fortuna influencia o palpite final dos participantes –, mas essa sucessão causal *não ocorre em virtude da relação semântica entre eles*, porque não há tal relação. Nesse sentido, as respostas dos participantes que foram influenciados pelo efeito de ancoragem são irracionais.

O título da obra de Kahneman, Rápido e Devagar, designa dois sistemas mentais que operam de maneira concorrente um com o outro[56]. Em geral, o sistema rápido está no controle das situações cotidianas, e o sistema devagar opera apenas quando realizamos um esforço consciente para resolver problemas. Dentre as funções do sistema rápido, encontram-se:

- detectar que um objeto está mais distante que outro;
- completar a frase 'pão com...';
- fazer cara de aversão ao ver uma foto horrível;
- detectar hostilidade no tom de voz de alguém;
- responder à pergunta '2 + 2 = ?';
- encontrar um movimento decisivo no xadrez (se você for um mestre enxadrista).

E dentre as funções do sistema devagar, temos:

- concentrar-se na voz de uma determinada pessoa em uma sala cheia e barulhenta;
- procurar uma mulher de cabelos brancos;
- monitorar a conveniência de seu comportamento numa situação social;
- contar as ocorrências da letra 'a' numa página de texto;
- preencher um formulário de imposto;

[56] Esses sistemas são substantivados pelo autor como um recurso didático. Porém, eles não correspondem a partes específicas do cérebro. O mais correto é pensar nos dois sistemas como dois tipos de *processos* mentais. Cada processo possui suas características específicas, seus pontos fortes e seus pontos fracos.

- verificar a validade de um argumento lógico complexo.

De acordo com Kahneman, o sistema rápido opera sem muito esforço, empregando mecanismos associativos de maneira automatizada, sobretudo nas situações em que precisamos agir rapidamente. O sistema devagar requer muito esforço para ser colocado em movimento, e só opera quando precisamos executar tarefas complicadas, que não permitem uma resolução automática e inconsciente. O ponto é que a psicologia contemporânea reconhece diferentes processos e mecanismos utilizados pelo cérebro, e não há a possibilidade de haver uma única teoria que dê conta de todos eles.

O objetivo desta Dissertação era apresentar a teoria computacional da mente como uma teoria a respeito de nossos processos racionais envolvendo atitudes proposicionais, reconstruindo os principais argumentos e motivações a favor da tese, e defendê-la dos ataques de Searle (1992). Esse objetivo foi ampliado após as críticas e sugestões feitas no Exame de Qualificação, e algumas alternativas teóricas à TCM foram brevemente consideradas no Capítulo 4, qual seja, o conexionismo e a cognição incorporada e situada. Para avaliar de maneira crítica essas alternativas, seriam necessários outros trabalhos e outros anos de estudo. Por essa razão, os modelos de processamento paralelo e a cognição incorporada e situada foram apresentados de maneira introdutória, ressaltando algumas discordâncias que elas possuem com relação ao computacionalismo clássico e evidenciando as mudanças de perspectiva que elas trouxeram para as ciências cognitivas. Entretanto, algumas considerações podem ser feitas após todas essas reflexões.

A TCM, do modo como a apresentamos desde o início do projeto, nunca foi e não pode ser por nós considerada uma teoria global da mente. Há vários aspectos de nossa vida mental sobre os quais ela é silenciosa. De fato, sua única, embora não pequena, contribuição é avançar uma proposta explicativa a respeito de como nosso cérebro se engaja em processos mentais que preservam a verdade de estados psicológicos para outros estados psicológicos. A teoria tenta desvendar a arquitetura mental que sustenta o raciocínio, mas reconhece a existência de muitos outros processos que são inexplicáveis dessa perspectiva computacional (ver citação que abre este capítulo). Provavelmente, ela diz respeito a uma parte dos processos lentos aos quais Kahneman se refere, mas tem pouco ou nada a dizer sobre os processos rápidos, por exemplo.

Uma vez que se reconhece a limitação da teoria, abre-se a possibilidade de ver as alternativas não como concorrentes diretas, mas como teorias complementares a respeito da cognição. Reconhecemos a existência de processos mentais associativos, e os modelos de PDP parecem apropriados para explicar esses fenômenos. Reconhecemos também o papel secundário

que o corpo teve nas ciências cognitivas até a década de 1980, e a necessidade de leva-lo em conta nas explicações do comportamento inteligente, bem como das relações que ele possui com o meio no qual se encontra. Em resumo, acreditamos ser mais provável que haja uma maior integração entre essas diferentes vertentes, ao invés do prognóstico pessimista (em certo sentido) de que uma irá suplantar a outra. Mas isso apenas o futuro poderá nos dizer.

REFERÊNCIAS

BAKER, Mark. **The Atoms of Language**. New York: Basic Books, 2001.

BARSALOU, Lawrence. Perceptual Symbol Systems. **Behavioral and Brain Sciences**, [*S. l.*], v. 22, p. 577-660, 1 jan. 1999.

_____. Situated conceptualization: theory and application. Separata de: COELLO, Yann; FISCHER, Martin. **Perceptual and Emotional Embodiment**: foundations of embodied cognition. London: Routledge, 2016. v. 1.

BARSALOU, Lawrence *et al.* Grounding conceptual knowledge in modality-specific systems. **TRENDS in Cognitive Sciences**, [*S. l.*], v. 7, p. 84-91, 2003.

BERMÚDEZ, José. **Philosophy of psychology**: a contemporary introduction. New York: Routledge, 2005.

_____. **Cognitive science**: a contemporary introduction. New York: Cambridge University Press, 2014.

BENNETT, Maxwell; DENNETT, Daniel; HACKER, Peter; SEARLE, John. **Neuroscience and philosophy**: brain, mind, and language. New York: Columbia University Press, 2007.

BIRABEN, Rodolfo. **Tese de Church**: algumas questões histórico-conceituais. 1994. Dissertação (Mestrado em Filosofia) - Universidade Estadual de Campinas, Instituto de Filosofia e Ciências Humanas, [*S. l.*], 1994.

_____. **Questões conceituais de computabilidade**. 2001. Tese (Doutorado em Filosofia) - Universidade Estadual de Campinas, Instituto de Filosofia e Ciências Humanas, [*S. l.*], 2001.

BLOCK, Ned. **Consciousness, function, and representation**: collected papers. Cambridge: The MIT Press, 2007.

BOCCARDI, Emiliano. Who's Driving the Syntactic Engine?. **Journal for General Philosophy of Science**, [*S. l.*], v. 40, p. 23-50, 2009.

BRADDON-MITCHELL, David; JACKSON, Frank. **Philosophy of Mind and Cognition**. Oxford: Blackwell, 1996.

BROENS, Mariana; GONZALEZ, Maria. Um estudo do conhecimento não proposicional no contexto da teoria da cognição incorporada e situada. **Manuscrito**, [*S. l.*], v. 29, p. 729-751, 2006.

BROOKS, Rodney. A Robust Layered Control System for a Mobile Robot. **IEEE Journal of Robotics and Automation**, [*S. l.*], v. 2, p. 14-23, mar. 1986.

BURGE, Tyler. **Origins of objectivity**. New York: Oxford University Press, 2010.

CHALMERS, David. **The character of consciousness**. New York: Oxford University Press, 2010.

CHEMERO, Anthony. **Radical Embodied Cognitive Science**. Cambridge: The MIT Press, 2009.

CHOMSKY, Noam. **Syntactic structures**. 2. ed. Berlim: Mounton de Gruyter, 2002.

CHURCHLAND, Paul. Eliminative materialism and the propositional attitudes. **Journal of Philosophy**, [*S. l.*], v. 78, n. 2, p. 67-90, 10 fev. 1981.

COPELAND, Jack. **Inteligencia artificial**: una introducción filosófica. Madrid: Alianza Editorial, 1993.

_____. What is computation?. **Synthese**, [*S. l.*], v. 108, p. 335-359, 10 set. 1996.

DAWSON, Michael. **Minds and Machines**: connectionism and psychological modeling. Oxford: Blackwell, 2004.

DENNETT, Daniel. **The intentional stance**. Cambridge: The MIT Press, 1989.

D'OTTAVIANO, Itala; FILHO, Ettore. Basic concepts of systemics. *In*: PEREIRA JR., Alfredo; PICKERING, William; GUDWIN, Ricardo. **Systems, self-organization and information**: an interdisciplinary perspective. New York: Routledge, 2018. p. 47-63.

ENDERTON, Herbert. **Elements of set theory**. New York: Academic Press, 1977.

FODOR, Jerry. **Psychological explanation**: an introduction to the philsophy of psychology. New York: Random House, 1968.

_____. **The Language of Thought**. Cambridge: Harvard University Press, 1975.

_____. The mind-body problem. **Scientific American**, [*S. l.*], v. 244, p. 114-123, 1 jan. 1981.

_____. **Psychosemantics**: the problem of meaning in the philosophy of mind. London: The MIT Press, 1987.

_____. **A Theory of Content and Other Essays**. London: The MIT Press, 1992.

_____. **The Mind Doesn't Work That Way**: the scope and limits of computational psyhchology. London: The MIT Press, 2001.

FREGE, Gottlob. **Lógica e Filosofia da Linguagem**. São Paulo: Edusp, 2009.

GARDNER, Howard. **The Mind's New Science**: a history of cognitive revolution. New York: Basic Books, 1985.

GLENBERG, Arthur; KASCHAK, Michael. Grounding language in action. **Psychonomic Bulletin & Review**, [*S. l.*], v. 9, p. 558-565, 2002.

GÖDEL, Kurt. On the completeness of the calculus of logic. Separata de: FEFERMAN, Solomon *et al*. **Kurt Gödel**: Collected works. Oxford: Oxford University Press, 1986. v. I, p. 61-101.

GORMAN, Paul; SEJNOWSKI, Terrence. Analysis of hidden units in a layered network trained to classify sonar targets. **Neural Networks**, [*S. l.*], v. 1, p. 75-89, 1988.

GRAHAM, George. Behaviorism. *In*: **Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2019. Disponível em: https://plato.stanford.edu/entries/behaviorism/. Acesso em: 29 jun. 2021.

HAACK, Susan. **Filosofia das lógicas**. São Paulo: Editora Unesp, 1998.

HEIL, John. **Philosophy of mind**: a contemporary introduction. 3. ed. New York: Routledge, 2013.

HINTON, George. Learning distributed representations of concepts. **Proceedings of the eighth annual conference of the Cognitive Science Society**, [*S. l.*], p. 1-12, 1986.

HUME, David. **An enquiry concerning human understanding**. Cambridge: Cambridge University Press, 2007.

HUTO, D; MYIN, E. **Radicalizing enactivism: basic minds without conent**. Cambridge, Massachusetts: MIT Press, 2013.

IACOBONI, Marco; DAPRETTO, Mirella. The mirror neuron system and the consequences of its dysfunction. **Nature reviews of neuroscience**, [*S. l.*], v. 7, p. 942-951, 2 dez. 2006.

IMMERMAN, Neil. Computability and complexity. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2018. Disponível em: https://plato.stanford.edu/entries/computability. Acesso em: 15 out. 2020.

JACKMAN, Henry. Meaning Holism. *In*: **Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2020. Disponível em: https://plato.stanford.edu/entries/meaning-holism/. Acesso em: 9 set. 2021.

JACKSON, Frank. Epiphenomenal qualia. **The Philosophical Quarterly**, [*S. l.*], v. 32, p. 127-136, out. 1982.

KAHNEMAN, Daniel. **Rápido e Devagar**: duas formas de pensar. Rio de Janeiro: Objetiva, 2011.

KAPLAN, Frederic; OUDEYER, Pierre-Yves; BERGEN, Benjamin. Computational models in the debate over language learnability. **Infant and Child Development**, [*S. l.*], v. 17, p. 55-80, 2008.

KREMPEL, Raquel. **An essay on the language of thought**. 2018. Tese (Doutorado em Filosofia) - Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, [*S. l.*], 2018.

KRIPKE, Saul. **Naming and Necessity**. Cambridge: Harvard University Press, 1972.

LAIRD, Johnson. Mental models in cognitive science. **Cognitive Science**, [*S. l.*], v. 4, p. 71-115, 1980.

LEVIN, Janet. Functionalism. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2018. Disponível em: https://plato.stanford.edu/entries/functionalism/. Acesso em: 16 fev. 2021.

LEVINE, Joseph. Materialism and qualia: the explanatory gap. **Pacific Philosophical Quarterly**, [*S. l.*], v. 64, p. 354-361, 1983.

LOWE, Edward. **An introduction to the philosophy of mind**. Cambridge: Cambridge University Press, 2000.

LYCAN, William. **Filosofia da Linguagem**: uma introdução contemporânea. Nova York: Routledge, 2000.

MAL'CEV, Anatoly. **Algorithms and recursive functions**. Wolters: Noordhof, 1970.

MARCUS, Gary. **The Algebraic Mind**: Integrating connectionism and cognitive science. Cambridge: The MIT Press, 2001.

MARR, David. **Vision**: a computational investigation into the human representation and processing of visual information. London: The MIT Press, 1982.

MILLER, George. The magical number 7, plus or minus two: some limits in our capacity for processing information. **The Psychological Review**, [*S. l.*], v. 63, n. 2, p. 81-97, mar 1956.

MINSKY, Marvin; PAPERT, Seymour. **Perceptrons**: an introduction to computational geometry. 3. ed. Cambridge: The MIT Press, 1969.

MORTARI, Cezar. **Introdução à lógica**. São Paulo: Editora Unesp, 2001.

MURPHY, Gregory. **The Big Book of Concepts**. Cambridge: The MIT Press, 2002.

MUSALL, Simon; URAI, Anne; SUSSILLO, David; CHURCHLAND, Anne. Harnessing behavioral diversity to understand neural computations for cognition. **Current Opinion in Neurobiology**, [*S. l.*], v. 58, p. 229-238, 2019.

NAGEL, Thomas. What is it like to be a bat?. **The Philosophical Review**, [*S. l.*], v. 83, p. 435-450, 1974.

OCKHAM, William of. **Summa logicae**: part I. London: University of Notre Dame Press, 1974.

PANTALEÃO, Nathália. **Os Limites da Teoria Computacional da Mente e o Papel do Corpo na Capacidade Semântica**. 2021. Tese (Doutorado em Filosofia) - Instituto de Filosofia e Ciências Humanas, Universidade Estadual de Campinas, [*S. l.*], 2021.

PEARL, Lisa. Using computational modeling in language acquisition research. *In*: BLOM, Elma; UNSWORTH, Sharon. **Experimental methods in language acquisition research**. Amsterdam: Cambridge University Press, 2010. p. 163-184.

PENROSE, Roger. **Shadows of the mind**: a search for the missing science of consciousness. New York: Oxford University Press, 1994.

PFEIFER, Rolf; BONGARD, Josh. **How the body shapes the way we think**: a new view of intelligence. Cambridge: The MIT Press, 2007.

PICCININI, Gualtiero. Physical Computation: a mechanistic account. Oxford: Oxford University Press, 2015.

_____. Computation in Physical Systems. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2017. Disponível em: https://plato.stanford.edu/entries/computation-physicalsystems. Acesso em: 19 out. 2020.

PICCININI, Gualtiero; MALEY, Corey. Computation in Physical Systems. In: Stanford Encyclopedia of Philosophy. [S. l.]: Edward N. Zalta, 2021. Disponível em: https://plato.stanford.edu/entries/computation-physicalsystems/. Acesso em: 8 set. 2021.

PINKER, Steven. **Language Learnability and Language Development**. London: Harvard University Press, 1984.

PINKER, Steven; PRINCE, Alan. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. **Cognition**, [*S. l.*], v. 28, p. 73-193, 1988.

PUTNAM, Hilary. Minds and machines. *In*: PUTNAM, Hilary. **Mind, language and reality**: philosophical papers. Cambridge: Cambridge University Press, 1975a. v. 2, cap. 18, p. 362-385.

_____. The nature of mental states. *In*: PUTNAM, Hilary. **Mind, language and reality**: philosophical papers. Cambridge: Cambridge University Press, 1975b. v. 2, cap. 20, p. 362-385.

_____. The meaning of 'meaning'. *In*: PUTNAM, Hilary. **Mind, Language and Reality**: philosophical papers. Cambridge: Cambridge University Press, 1975c. cap. 12, p. 215-271.

PYLYSHYN, Zenon. **Computation and cognition**: toward a foundation for cognitive science. London: The MIT Press, 1984.

RESCORLA, Michael. Levels of computational explanation. *In*: POWERS, Thomas. **Philosophy and computing**. [*S. l.*: *s. n.*], 2017a. v. 128, p. 5 - 28.

_____. From Ockham to Turing – and back again. *In*: FLOYD, Juliet; BOKULICH, Alisa. **Philosophical explorations of the legacy of Alan Turing**. [*S. l.*]: Springer International Publisher, 2017b. p. 279-304. ISBN 978-3-319-53280-6.

_____. The language of thought hypothesis. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2020. Disponível em: https://plato.stanford.edu/entries/language-thought. Acesso em: 15 out. 2020.

_____. The computational theory of mind. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2020. Disponível em: https://plato.stanford.edu/entries/computational-mind. Acesso em: 15 out. 2020.

REDINGTON, Martin; CHATER, Nick; FINCH, Steven. Distributional information: a powerful cue for acquiring syntactical categories. **Cognitive Science**, [*S. l.*], v. 22, p. 425-469, 1998.

RIEL, Raphael; GULICK, Robert. Scientific Reduction. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2019. Disponível em: https://plato.stanford.edu/entries/scientific-reduction/. Acesso em: 11 fev. 2021.

ROSENBLATT, Frank. The perceptron: A probabilistic model for information storage and organization in the brain. **The Psychological Review**, [*S. l.*], v. 65, p. 386-408, 1958.

RUMELHART, David; MCCLELLAND, James. **Parallel Distributed Processing**: explorations in the microstructure of cognition. Cambridge: The MIT Press, 1986.

SEARLE, John. Minds, brains, and programs. **The Behavioral and Brain Sciences**, [*S. l.*], v. 3, p. 417-457, 1980.

_____. **The rediscovery of the mind**. Cambridge: The MIT Press, 1992.

SHAGRIR, Oron. Why we view the brain as a computer. **Synthese**, [*S. l.*], v. 153, n. 3, p. 393-416, 23 nov. 2006.

SHANNON, Claude. A mathematical theory of communication. **The Bell System Technical Journal**, [*S. l.*], v. 27, p. 379-423, jul. 1948.

_____. A universal Turing Machine with two internal states. *In*: WYNER, Aaron. **Claude E. Shannon: Collected Papers**. [*S. l.*]: Wiley-IEEE Press, 1993. p. 733-741.

SHAGRIR, Oron. Why we view the brain as a computer. **Synthese**, [*S. l.*], v. 153, n. 3, p. 393-416, 23 nov. 2006.

_____. In defense of the semantic view of computation. **Synthese**, [*S. l.*], v. 197, p. 4083-4108, 11 out. 2018.

SHAPIRO, Lawrence; SPAULDING, Shannon. Embodied Cognition. *In*: **Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2021. Disponível em: https://plato.stanford.edu/entries/embodied-cognition/#Bib. Acesso em: 12 ago. 2021.

SHEPARD, Roger; METZLER, Jacqueline. Mental rotation of three dimensional objects. **Science**, [*S. l.*], v. 171, p. 701-703, 19 fev. 1971.

SKINNER, Burrhus. **Science and human behavior**. New York: Macmillan, 1953.

SKLAR, Lawrence. **Physics and chance**: philosophical issues in the foundation of statistical mechanics. Cambridge: Cambridge University Press, 1993.

SMITH, Peter. **An introduction to Gödel's Theorem**. New York: Cambridge University Press, 2013.

STOLJAR, Daniel. Physicalism. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2017. Disponível em: https://plato.stanford.edu/entries/physicalism/. Acesso em: 15 out. 2020.

SZABÓ, Zoltán. Compositionality. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2020. Disponível em: https://plato.stanford.edu/entries/compositionality/. Acesso em: 21 fev. 2021.

THAGARD, Paul. Cognitive science. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2019. Disponível em: https://plato.stanford.edu/entries/cognitive-science/. Acesso em: 15 out. 2020.

TOLMAN, Edward. Cognitive maps in rats and men. **The Psychological Review**, [*S. l.*], v. 55, p. 189-208, 1948.

TUCKER, Mike; ELLIS, Rob. On the Relations Between Seen Objects and Components of Potential Actions. **Journal of Experimental Psychology**, [*S. l.*], v. 24, n. 3, p. 830-846, jul. 1988.

TURING, Alan. On computable numbers: with an application to the Entscheidungsproblem. **Proceedings of the London Mathematical Society**, [*S. l.*], v. 42, p. 230-265, 1 nov. 1936.

_____. Computing machinery and intelligence. **Mind**, [*S. l.*], v. 49, p. 433-460, 1 out. 1950.

TYE, Michael. Qualia. *In*: **The Stanford Encyclopedia of Philosophy**. [*S. l.*]: Edward N. Zalta, 2018. Disponível em: https://plato.stanford.edu/entries/qualia. Acesso em: 15 out. 2020.

WATSON, John. **Psychology from the standpoint of a behaviorist**. Philadelphia: Lippincot, 1919.

WEBB, Barbara. A Robot Cricket. **Scientific American**, [*S. l.*], v. 275, p. 94-99, 1 jan. 1996.

WHITING, Daniel. Conceptual Role Semantics. *In*: **Internet Encyclopedia of Philosophy**. [*S. l.*]: James Fieser, sem data. Disponível em: https://iep.utm.edu/conc-rol/. Acesso em: 11 fev. 2021.

APÊNDICE

Este Apêndice pretende ilustrar, através do exemplo a seguir, a ideia apresentada no Capítulo 1 de que a compreensão do funcionamento do cérebro e seus muitos sistemas envolve vários níveis teóricos de abstração. Em especial, estamos interessados no caráter funcional de explicações do nível da psicologia cognitiva, as quais se abstraem de detalhes fisiológicos a respeito da microestrutura dos neurônios em favor de uma caracterização mais abrangente do seu comportamento. Outro ponto a ser ilustrado é o apresentado no Capítulo 3, sobre o caráter informacional dos processos cognitivos, mesmo inconscientes. Assim, expomos uma explicação esquemática desse fenômeno visual.

Algumas ilusões ópticas são vantajosas de um ponto de vista adaptativo. Esse é o caso das Bandas de Mach (Figura A-1), uma ilusão na qual percebemos de maneira mais aguda variações na intensidade de luz entre superfícies claras e escuras. Faixas mais à direita são mais claras que faixas à esquerda. Porém, percebemos também uma variação de intensidade mais sutil, interna a cada faixa, como se dentro de cada uma delas o canto direito fosse levemente mais escuro que o canto esquerdo. Essa variação interna não existe, é uma ilusão gerada pelo modo como as células nervosas da retina funcionam.

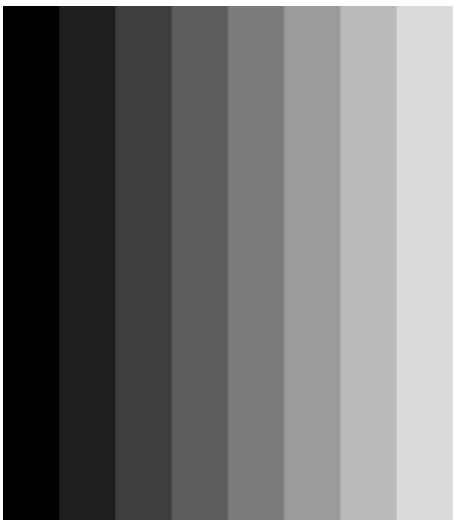

Figura A-1 - Bandas de Mach. Se olharmos atentamente para cada faixa de cor, veremos uma variação de intensidade dentro da própria faixa. Essa variação não existe.

No entanto, essa ilusão cumpre uma função interessante em nosso aparato visual. Ela é gerada pelo nosso mecanismo detector de contornos, que facilita a identificação visual das bordas dos objetos. A neuroanatomia do olho é representada na Figura A-2 abaixo. As células receptoras de luz (também chamadas de *transdutoras*, pois transformam sinais luminosos em sinais elétricos) são compostas, basicamente, de cones e bastonetes. Cones e bastonetes cumprem funções

bastante diferentes, mas, grosso modo, cones fornecem informações sobre cores e os bastonetes sobre intensidade de luz. A figura ilustra também a organização neural das células que transmitem ao cérebro as informações captadas pelos receptores de luz, compondo uma complexa rede neural.

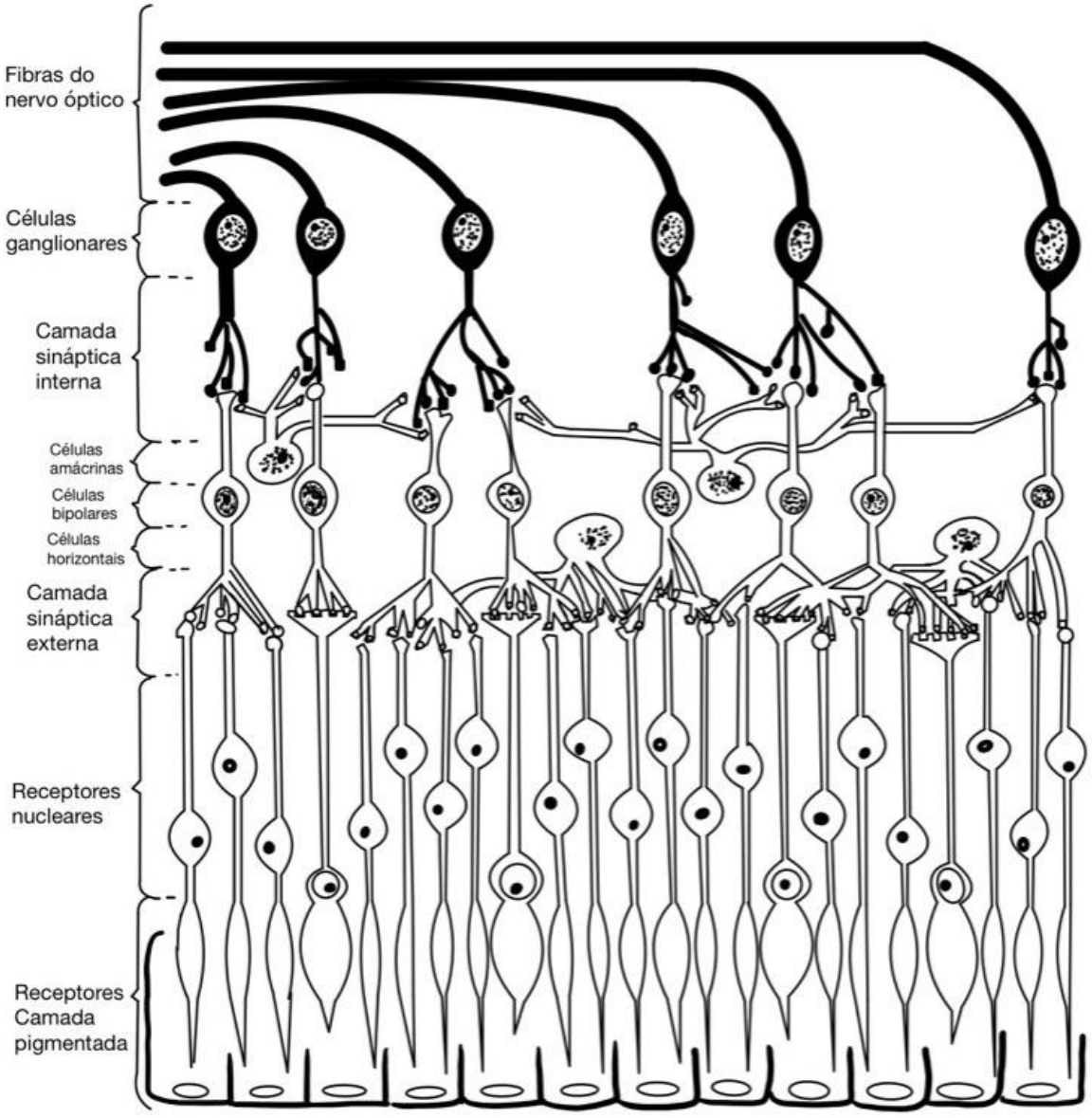

Figura A-2 - Neuroanatomia do olho (adaptado de Lindsay & Norman 1977).

Um neurônio é composto de um corpo celular (que recebe impulsos elétricos), um axônio ou fibra nervosa (estrutura prolongada que serve como condutor elétrico) e os dendritos (ramificações que enviam a outros neurônios sinais elétricos). As sinapses são as regiões que

conectam o corpo celular de um neurônio com dendritos de outros neurônios. No caso da retina, a organização da rede pode ser dividida em duas. Algumas células nervosas são responsáveis por enviar os sinais dos receptores de luz aos centros de processamento mais altos do cérebro, compondo uma *malha vertical*. As células bipolares e ganglionares executam essa função (o nervo óptico é composto dos axônios das células ganglionares, e partem da parte de trás do olho até o córtex visual primário, localizado na parte de traz da cabeça). Outras células fazem conexões entre as células da rede vertical, compondo uma *malha horizontal*. Essas são as células horizontais e as células amácrinas. O efeito das Bandas de Mach se deve a essa estrutura horizontal de neurônios. O que ocorre é que, devido a essas conexões horizontais, os sinais recebidos em um receptor influenciam os sinais de receptores vizinhos.

Abstraímo-nos, porém, dessas distinções sutis entre os tipos de células que compõem os receptores e os tipos de neurônios que compõem a organização horizontal de processamento da retina. Embora discriminações funcionais possam ser feitas a respeito dos cones e bastonetes, estamos interessados apenas no fato de que eles transformam inputs luminosos em outputs elétricos. Portanto, tratamos apenas de *receptores*. O mesmo se aplica às células horizontais e às células amácrinas. No caso, estamos interessados apenas em como elas influenciam os sinais das células vizinhas. Portanto, chamamos as duas simplesmente de *células neurais*. Receptores são representados como na Figura A-3.

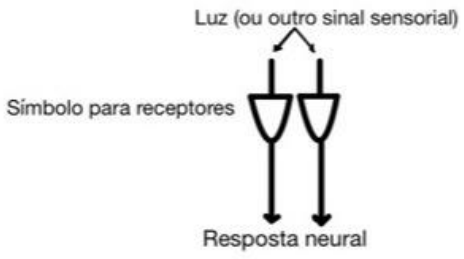

Figura A-3 – Representação dos receptores (adaptado de Lindsay & Norman (1977).

Estímulos luminosos serão representados por um número. De maneira simplificada, um estímulo luminoso de intensidade 10 indica que o receptor gerará uma resposta neural de valor 10. Assim, um receptor recebe um estímulo luminoso (com intensidade determinada) e responde com um sinal elétrico que serve de entrada para uma célula neural. Ignoramos os mecanismos através dos quais um estímulo físico é transformado em sinal elétrico. Mas isso não implica em prejuízo na explicação.

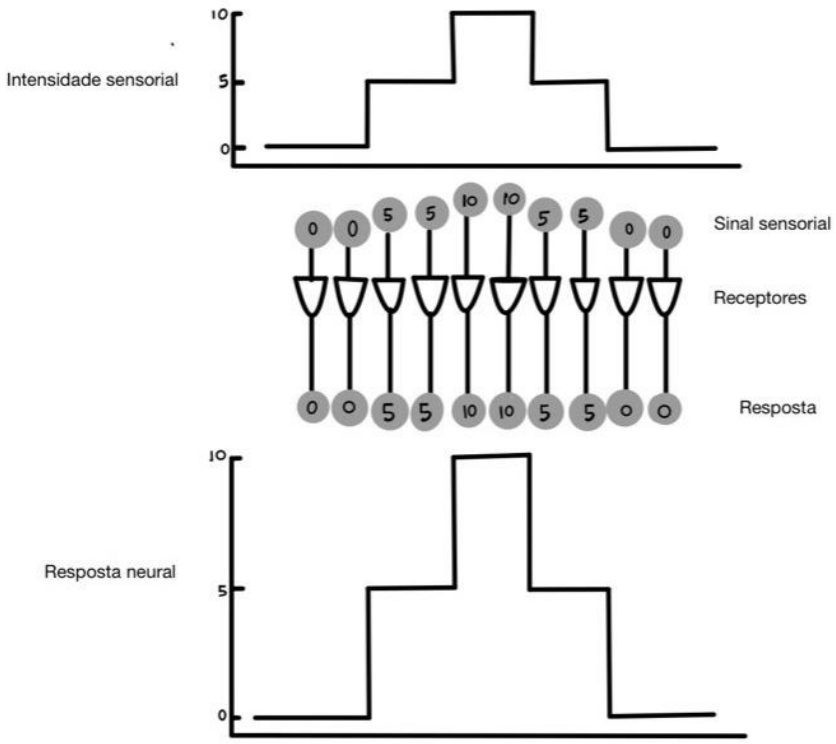

Figura A-4 – Representações gráficas de entrada e de saída dos receptores (adaptado de Lindsay & Norman (1977)).

As células neurais são representadas por um círculo, como na Figura A-5.

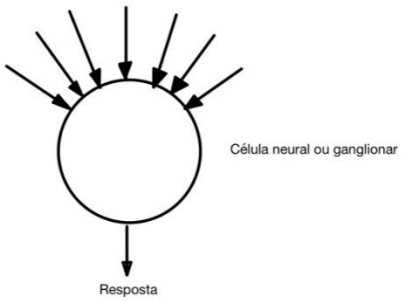

Figura A-5 – Esquema de célula neural (adaptado de Lindsay & Norman (1977)).

Células neurais se conectam através de dois tipos de sinais: excitatórios e inibitórios. Conexões excitatórias tem um valor positivo associado, e conexões inibitórias um valor negativo associado, os quais influenciarão a saída da célula. Além disso, células neurais possuem um nível de disparo espontâneo (taxa de fundo), que ocorre mesmo quando a célula está inativa. Suponha, assim, que uma célula neural possui uma taxa de fundo igual a 100 e recebe um input de valor 80, com fator excitatório de 0,25. Seu output será o valor da taxa de fundo somado ao input multiplicado pelo fator excitatório: $100 + (80 \times 0{,}25) = 120$.

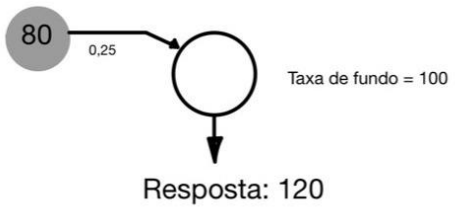

Figura A-6 – Esquema de célula neural (adaptado de Lindsay & Norman (1977)).

O mesmo vale para fatores inibitórios, com a diferença de que a multiplicação entre input e fator inibitório será subtraída da taxa de fundo (ao invés de somada). Uma célula neural com taxa de base igual a 100, input de valor 80, com fator inibitório de 0,25 terá um output de 100 – (80 × 0,25) = 80.

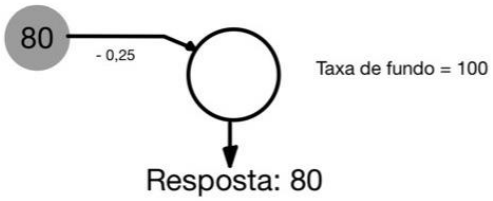

Figura A-7 – Esquema de célula neural (adaptado de Lindsay & Norman (1977)).

Como os contornos de uma cena são apreendidos a partir de variações na intensidade de luz, é importante que nosso sistema visual seja bastante sensível a essas variações. Sem a malha horizontal de processamento, os sinais captados pelos receptores seriam enviados sem nenhum tipo de modulação, de modo que a mensagem enviada ao cérebro constituiria uma representação fidedigna dos padrões luminosos interceptados pela retina.

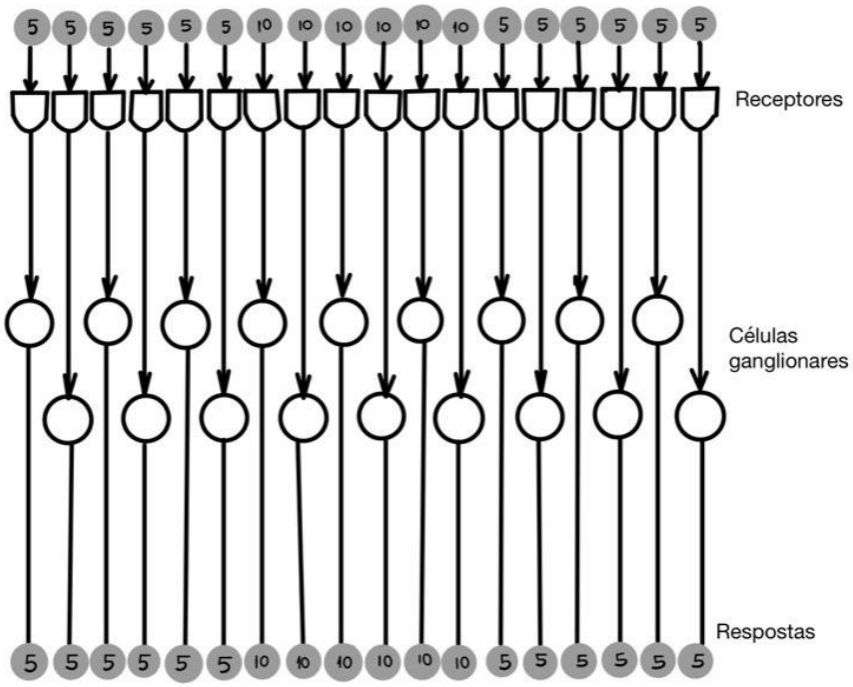

Figura A-8 - Representação unidimensional de uma sequência de receptores e células ganglionares, sem processamento horizontal. As saídas são idênticas às entradas.

Mas, uma vez que o problema de identificar bordas consiste no problema de identificar variações bruscas na intensidade de luz, é útil que essas variações sejam "realçadas" pelo próprio sistema. Pois, se uma variação tênue na intensidade de luz corresponder a uma borda no ambiente externo, poderia haver incertezas na medição, associadas a ruídos do próprio sistema, por exemplo. Em suma, o que nós queremos é que a rede de neurônios permaneça constante quando não houver variações na intensidade de luz, e que ela forneça respostas vigorosas a variações nos padrões luminosos.

Imagem A-1 - Límulo ou caranguejo-ferradura.

Grande parte do conhecimento do funcionamento do olho se deve aos caranguejos-ferradura (Imagem A-1). Pesquisadores isolaram seus receptores visuais para estimulá-los separadamente. Embora seja impossível um receptor A excitar um receptor B vizinho, quando os dois são estimulados simultaneamente é possível que o receptor A iniba ou enfraqueça os

sinais enviados pelo receptor B às células ganglionares. Esse mecanismo é chamado de *inibição lateral*, que consiste na operação de uma célula modificar a operação em curso de uma célula da vizinhança. Isso é feito através do processamento horizontal apresentado acima, pelos mecanismos de sinais excitatórios e inibitórios.

A Figura A-9 ilustra uma fileira unidimensional de receptores e células neurais, onde há uma conexão horizontal entre os neurônios com fator inibitório de -2. Cada neurônio se conecta não com seu vizinho imediato, mas com duas células que se encontram duas posições para trás e a outra duas posições para frente. Os pontos críticos são onde a intensidade de luz varia de 5 para 10 e de 10 para 5 novamente. A malha horizontal de processamento acentua as variações de luz ao mesmo tempo em que preserva informações a respeito da intensidade relativa entre uma superfície e outra e mantém inalterável a intensidade luminosa onde ela é de fato constante.

Podemos ver agora como a ilusão das Bandas de Mach é gerada. Quando a luz varia de intensidade, essa variação é acentuada pelo próprio sistema de processamento horizontal. De que modo ocorre essa acentuação? Sempre que superfícies claras são colocadas ao lado de superfícies escuras, nossa percepção é de que a superfície clara é mais clara do que realmente é (na região em que se encontra com a superfície escura), e também percebemos a superfície escura mais escura do que realmente é (na região em que se encontra com a superfície clara). Nossa percepção das Bandas é representada pela Figura A-10.

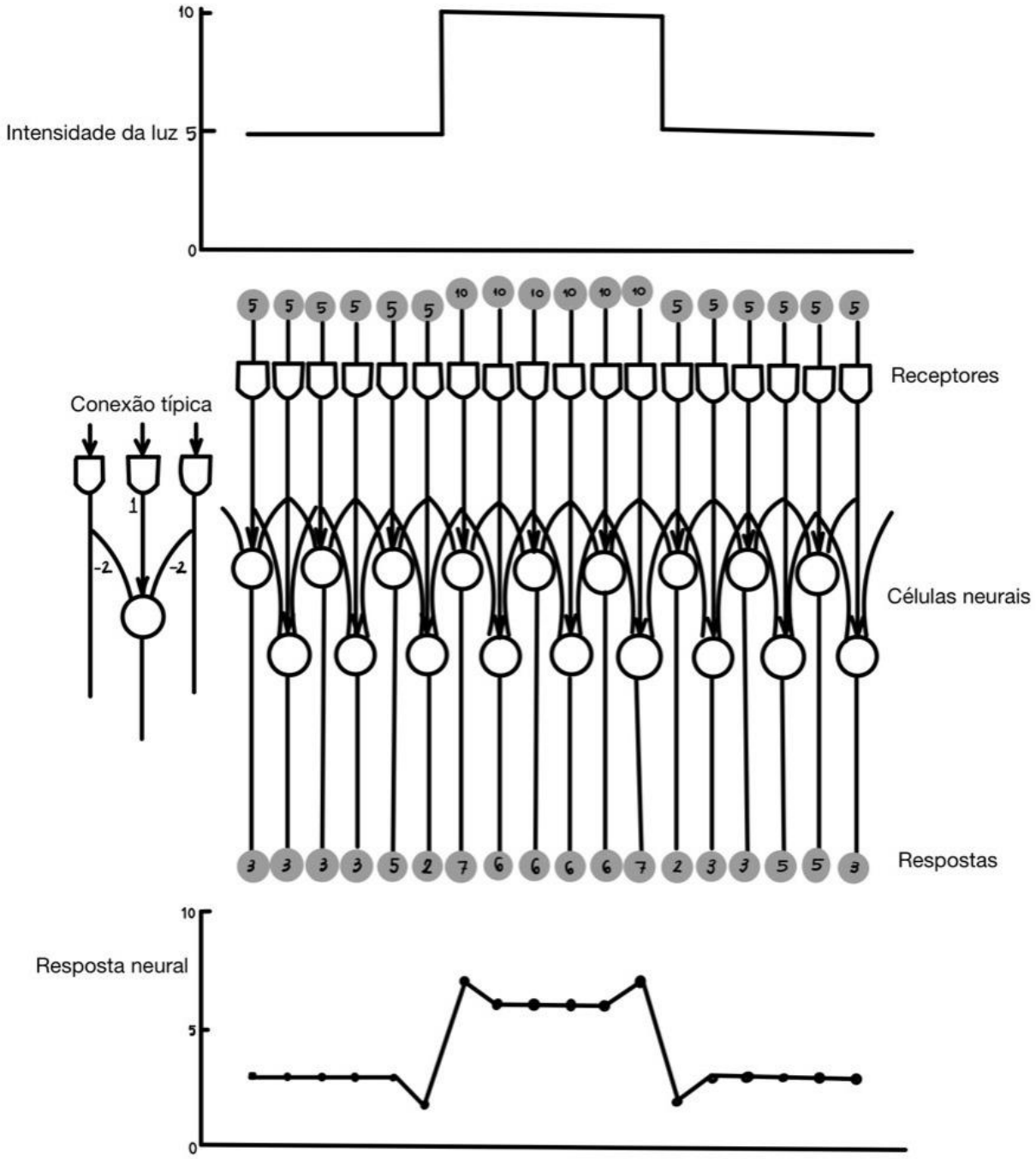

Figura A-9 - Representação unidimensional de uma sequência de receptores e células ganglionares com processamento horizontal. A saída acentua os limites identificados na entrada.

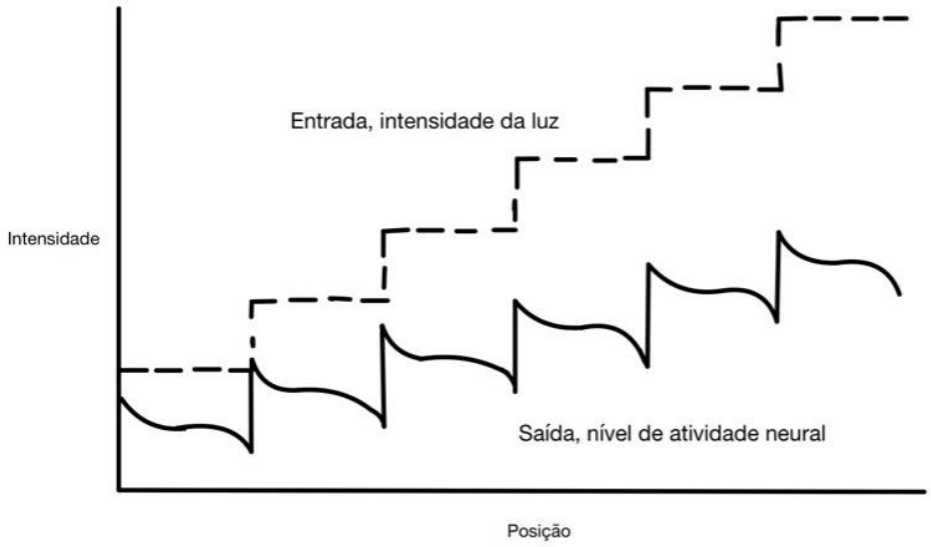

Figura A-10 – Nossa percepção das Bandas de Mach (adaptado de Lindsay & Norman (1977)).

Esse caso ilustra pontos importantes sobre como as investigações são conduzidas nas ciências cognitivas. Talvez o mais importante deles seja a respeito do nível de abstração em que a explicação é construída. A noção de que a compreensão do sistema nervoso admite diversos níveis de explicação foi abordada no Capítulo 1, mas sentimos a necessidade de que algum exemplo concreto fosse apresentado. Nosso objetivo aqui é o de que fique evidente o nível intermediário de caracterização do funcionamento de nosso sistema visual — não tão abstrato quanto o nível da psicologia de senso comum e nem tão básico quanto o nível de caracterização biológico[57]. Detalhes a respeito da microestrutura e dos processos químicos responsáveis pelo funcionamento neural são omitidos em favor de uma caracterização mais geral dos fenômenos observados, privilegiando uma descrição funcional das partes que compõem o sistema. Em razão disso, somos capazes de observar congruências nos sistemas visuais de organismos completamente diferentes[58]. Por outro lado, a explicação fornecida reside num nível de abstração mais baixo que o nível da psicologia de senso comum, pois, como vimos, a explicação funcional do mecanismo de inibição lateral está diretamente ancorada na caracterização neurofisiológica da retina.

Outro ponto importante é o caráter informacional da explicação. Naturalmente vemos os sinais elétricos enviados pelo nervo óptico como carregando informação a respeito de propriedades geométricas dos objetos do campo de visão. É quase inevitável descrever os

[57] Não esperamos, com isso, implicar que existem três níveis de caracterização.

[58] Embora, como foi notado anteriormente, grande parte desse conhecimento tenha sido obtido a partir de análises do sistema visual de caranguejos-ferradura, Lindsay & Norman observam "In mammals, the retinal operations differ somewhat from those of a crab, but the basic features of the analysis are similar" (1977, p. 215).

processos dessa forma. Se nossos estados perceptuais conscientes carregam informação a respeito dos objetos externos, por que não dizer que as respostas elétricas fornecidas pelos receptores carregam alguma informação, já que esse é o principal canal através do qual extraímos a informação visual? Além disso, de uma perspectiva informacional conseguimos compreender como as células horizontais modulam os sinais, tornando o sistema todo mais eficiente enquanto mecanismo detector de bordas. Se adotarmos esse tipo de abordagem (informacional), podemos descrever os processos como mentais, em oposição a unicamente fisiológicos, já que os estados neurofisiológicos podem ser considerados como intencionais.

Não estamos dizendo que essa é a única interpretação dos processos, mas que é uma interpretação natural de como os mecanismos cerebrais podem ser analisados.