CDD: 001.535

O FÍSICO E O MENTAL: INTELIGÊNCIA ARTIFICIAL E O PROBLEMA MENTE-CÉREBRO

JOÃO DE FERNANDES TEIXEIRA

Universidade Federal de São Carlos
São Carlos, SP.

O artigo trata do problema filosófico das relações mente-corpo e discute as possíveis contribuições da Inteligência Artificial e do modelo computacional da mente para a solução deste problema. A primeira parte retraça as raízes históricas da distinção entre o físico e o mental remetendo-nos aos trabalhos de R. Descartes. A segunda parte aborda algumas teorias contemporâneas tais como a análise linguística proposta por G. Ryle, as teorias da identidade mente-cérebro e o materialismo eliminativo. A terceira parte analisa o trabalho de H. Putnam e sua analogia entre mentes e máquinas. Finalmente, na última parte, abordamos novas metáforas computacionais baseadas em modelos conexionistas e discutimos em que sentido estas podem contribuir para esclarecer as dificuldades conceituais envolvidas no problema das relações mente-cérebro.

The paper focuses on the mind-body problem and discusses to what extent the computational model of the mind can cast some light on this traditional issue. Section I is devoted to trace the contours of the distinction between the mental and the physical by concentrating the analysis in the philosophical works of R. Descartes. Section II explores some contemporary approaches to this question such as the linguistic survey developed by G. Ryle, the identity-theory and eliminative materialism. Section III focuses on H. Putnam's functionalism and his analogy between minds and computers. Finally, in section IV, a new computer metaphor based on connectionism is proposed as an alternative to accommodate conceptual difficulties involved in the mind-body problem.

O surgimento do chamado modelo computacional da mente, formulado há algumas décadas, trouxe como conseqüência uma forte aproximação entre Inteligência Artificial e problemas filosóficos, tornando a chamada Filosofia da Mente em uma área de entrecruzamento de interesses comuns partilhados por disciplinas aparentemente tão díspares

como a Ciência da Computação e a Filosofia. Com efeito, a Inteligência Artificial abriu um leque de novas perspectivas para o tratamento de problemas filosóficos tradicionais realizando uma espécie de triangulação progressiva entre Filosofia, Psicologia e Ciência da Computação. O conceito de simulação cognitiva invadiu os domínios tradicionais da reflexão filosófica: progressivamente, os teóricos da Inteligência Artificial ou da Ciência Cognitiva passaram a desenvolver e implementar modelos computacionais com o objetivo de reformular e talvez solucionar problemas filosóficos tradicionais tais como as discussões que cercam a ontologia das nossas representações mentais ou o problema da natureza da intencionalidade e toda uma outra gama de problemas convencionais. Modelos computacionais para simular comportamento intencional, contextos opacos ou intensionais (com s) escritos em linguagem LISP passam cada vez mais a invadir os redutos sagrados da análise filosófica. O grande projeto da Inteligência Artificial – ou pelo menos o seu grande projeto filosófico – parece impor-se cada vez mais: trata-se de elaborar teorias da cognição que sejam empiricamente testáveis a partir de implementações práticas em computador reforçando de maneira cada vez mais acentuada os contornos de uma epistemologia naturalizada como fora protagonizada de forma programática por pensadores como Quine.

Nosso objetivo neste artigo é focalizar um destes interessantes entrecruzamentos entre Filosofia e Inteligência Artificial sem entretanto endossar as teses da simulação cognitiva que certamente forçar-nos-iam a discutir o estatuto científico e a validade das teorias propostas em Inteligência Artificial – uma questão que a nosso ver teria de ser abordada primariamente nos termos de uma discussão da testabilidade das teorias cognitivas. Nosso objetivo é um pouco mais modesto e ilustrativo: trata-se de mostrar como um problema filosófico, qual seja, o das relações entre o físico e o mental, ou entre mente e cérebro pode efetivamente ser tratado sob a ótica da Inteligência Artificial e como as contribuições desta, embora não trazendo soluções, colaboram no

sentido de reformulá-lo num esforço para que tal problema possa pelo menos ser enunciado de maneira adequada. Uma reformulação significará, antes de mais nada, a proposta de um modelo ou de uma metáfora conveniente que permita estabelecer uma relação inteligível entre o físico e o mental bem como uma possível comensurabilidade entre as disciplinas que os abordam (por exemplo, a Neurofisiologia e a Psicologia, respectivamente).

Estabelecer metáforas, algumas delas com elevado valor heurístico, tem sido a quase totalidade das atividades nas Ciências Cognitivas, a começar pela própria proposta do modelo computacional da mente e é neste tipo de tarefa que surgem interessantes parentescos entre a Inteligência Artificial, a Filosofia da Ciência, a Filosofia da Linguagem e a própria filosofia analítica entendida como análise da linguagem. Estabelecer uma teoria científica como sendo uma metáfora significa também reconhecer os limites e o escopo de nossa própria atividade cognitiva. No caso do problema das relações mente-cérebro estabelecer uma teoria como uma metáfora ou um modelo significa evitar a ilusão metafísica pré-kantiana de que haveria um lugar privilegiado ou transcendental do qual poderíamos decidir acerca da natureza da nossa própria atividade mental – um lugar que nos situaria para além dos limites do dizível e do concebível e de onde a mente poderia decidir, através de atos de cognição, acerca de sua própria natureza última.

Começaremos rastreando nosso problema na história da Filosofia abordando em primeiro lugar alguns aspectos da obra de Descartes. Estes comentários preliminares servirão como uma primeira aproximação dos contornos do problema que desejamos examinar e ocuparão a primeira parte deste trabalho. Num segundo momento examinaremos versões conteporâneas do problema mente-corpo com especial ênfase nas propostas materialistas que surgem no século XX. Nossa preocupação nestes dois estágios do nosso trabalho, será estabelecer os delineamentos do problema que desejamos tratar, deixando os detalhes da análise historiográfica para um segundo plano. Será a partir des-

tes delineamentos que teremos parâmetros para estabelecer um modelo que acomode as dificuldades teóricas que surgem do exame das relações entre o físico e o mental.

A terceira parte aborda uma contribuição importante da Inteligência Artificial no sentido de reformular o problema mente-corpo: o trabalho de H. Putnam e sua proposta materialista expressa na teoria da identidade teorética entre estados mentais e estados cerebrais. Mais do que um objetivo subsidiário ao nosso tema, a análise e crítica do modelo proposto por Putnam visa mostrar que a elaboração de uma metáfora adequada para conceber as relações mente-corpo passa, quase que necessariamente, por uma discussão linguística mas certamente não se resume a ela. Finalmente, na última parte, introduziremos os chamados modelos conexionistas ou modelos baseados em redes neurais e arquiteturas não-convencionais em Inteligência Artificial. Estes modelos constituirão um ponto de partida privilegiado para a elaboração de nossa proposta teórica no que diz respeito ao problema das relações mente-cérebro, fornecendo os elementos necessários para restabelecer uma unidade teórica coerente que permita superar algumas das dificuldades que apontamos nas seções anteriores. Sustentaremos que os modelos conexionistas constituem por si só a metáfora que estamos buscando; uma metáfora que por suas fortes conotações imagéticas tem sido sistemática e indevidamente confundida com o objeto que ela visa modelar, qual seja, a atividade mental. Tomar o cérebro como ponto de partida para conceber a atividade mental não significa, como querem alguns, afirmar que esta última pode legitimamente ser reduzida a uma específica arquitetura neuronal. Confundir modelos explicativos com o próprio objeto que desejamos explicar constitui uma falácia cujas raízes se encontram na interpretação realista da natureza das teorias científicas da mente – uma falácia que visa restabelecer o lugar privilegiado a partir do qual teríamos uma descrição unívoca e definitiva da própria natureza do mental.

I

O problema mente-corpo ou mente-cérebro é um dos mais tradicionais da filosofia moderna. O surgimento da filosofia cartesiana, na época clássica, constitui um marco importante na história desta questão. Com efeito, deve-se a Descartes a revolução filosófica que trouxe como conseqüência fundamental a separação entre mente e corpo, ou entre *res cogitans* e *res extensa*. Mas ao mesmo tempo em que efetuava esta mudança paradigmática na história da Filosofia, o cartesianismo deixava como herança não apenas o problema de formular uma articulação entre as duas substâncias de modo a poder explicar o comportamento, como também a questão da própria possibilidade de cognição do chamado mundo exterior – um problema que de modo algum é alheio aos teóricos da Inteligência Artificial.

A distinção cartesiana trouxe como conseqüência indesejável a necessidade de postular a existência de um Deus Todo-Poderoso que asseguraria a correspondência entre o mundo e a representação mental que dele formamos – a garantia metafísica de que o conhecimento poderia ser alcançado. A distinção cartesiana trouxe também a necessidade de atribuir estranhas propriedades a um órgão chamado "glândula pineal" que caracterizaria uma região híbrida a meio caminho entre mente e corpo mas possibilitaria sua articulação.

A má história da Filosofia tende a ver o cartesianismo como um sistema obsoleto – um mausoléu do dualismo que já teria cumprido seu papel histórico e que agora precisa ir para o olvido, juntamente com o problema das relações mente-corpo. Infelizmente não é o cartesianismo que é obsoleto ou equivocado e sim a interpretação que dele se fez como se de sua metafísica se pudesse derivar a separação entre duas *substâncias*.

A leitura habitual de Descartes parece basear-se numa confusão entre *separação* e *separabilidade* das substâncias pensante e material (ou extensa). E esta não é uma mera questão terminológica: sustentar a

separação entre as substâncias significa postular a existência de uma *substância pensante* que nos remeteria para uma inteligência pura ou uma realidade imaterial. Por outro lado, se sustentamos unicamente a separabilidade das substâncias, não postulamos nenhum domínio ontológico para além do extenso (e sensível) mas tão somente a necessidade de se estabelecer uma distinção categorial. O físico e o mental seriam separáveis, mas sua manifestação seria sempre conjunta e tudo se passaria como se estivéssemos estabelecendo uma distinção que seria muito mais metodológica do que propriamente ontológica[1]. Se seguirmos a primeira interpretação, rapidamente encontraremos a dificuldade, quase que insuperável, de explicar como a "substância imaterial" poderia interagir com o mundo da matéria, uma dificuldade que é acentuada pelo fato de que não estamos dispostos a abrir mão da idéia de que nossos comportamentos ou nossas ações seguem-se como resultado daquilo que ocorre na nossa mente. Se seguirmos a segunda interpretação, esta dificuldade seria superada, embora encontremos outras: seria praticamente impossível sustentar a imortalidade da alma após a morte, o que para Descartes parece ter sido uma tese metafísica de razoável importância[2]. Acrescenta-se a este problema outros que envolvem questões históricas e da coerência na interpretação da estrutura da própria obra cartesiana: se só existe separabilidade e não separação, como conceber a seqüência das *Meditações*? Em outras palavras, se o *Cogito* nos remetesse a um eu concreto, onde alma e corpo estivessem juntos, como seria possível o encadeamento das razões após a introdução da dúvida radical? Não desapareceria o eu juntamente com

[1] Este ponto de vista é sustentado por Alanen, L. (1981) e Teixeira, L. (1990).

[2] Veja-se por exemplo a seguinte passagem de Descartes: "Le corps, pris en général, est une substance, c'est pourquoi il ne périt point, mais le corps humain, en tant qu'il différe des autres corps n'est formé et composé que d'une certaine configuration de membres et d'autres semblables accidents; et lâme humaine, au contraire, n'est point ainsi composée d'accidents, mais est une pure substance... D'ou il s'ensuit que le corps humain peut facilement perir, mais que l'esprit ou lame de l'homme (ce que je ne distingue point) est immortelle de sa nature" (*Abregé des Méditations*, VII, pp. 13-14).

esta última, na medida em que este eu seria uma coisa do mundo? E se não há efetiva separação entre as duas substâncias, em que Descartes estaria se distinguindo de seus antecessores medievais?

Estas dificuldades de interpretação histórica e estrutural da obra de Descartes parecem ter contribuído definitivamente para que a primeira leitura tenha prevalecido na filosofia pós-cartesiana. Contudo, os movimentos filosóficos que sucederam a revolução cartesiana parecem ter se ressentido da incapacidade de formular uma solução para o problema que apontamos acima: como poderia a substância imaterial interagir com o mundo material? A dificuldade se acentua na medida em que não estamos dispostos a abrir mão da idéia habitual de que nossas ações seguem-se como resultado daquilo que ocorre na nossa mente. Para Descartes seria preciso encontrar algo que possibilitasse uma conexão entre as duas substâncias – uma conexão que deveria preservar a universalidade do princípio de causalidade e com ela a proposta de uma visão mecanicista do universo. A solução proposta por Descartes, qual seja, a doutrina da glândula pineal, notabilizou-se pela sua incapacidade de resolver o problema: a insatisfação com esta teoria parece ter sido a responsável pelo surgimento de teorias como o ocasionalismo de Malebranche e a teoria metafísica da harmonia pré-estabelecida proposta por Leibnitz. Uma alternativa às teorias da interação entre as duas substâncias seria sua unificação: este tipo de solução encontramos em teorias como o materialismo hobbesiano e numa série de outras tentativas que imprimiram à história da Filosofia um movimento pendular que se mantém até hoje, oscilando ora para posições marcadamente materialistas ora para posições exageradamente idealistas.

Os obstáculos teóricos oriundos das tentativas interacionistas constituem uma boa razão para que nos inclinemos em direção à segunda leitura de Descartes e sustentemos a separabilidade das substâncias. De acordo com este ponto de vista, falar de uma separação entre duas substâncias ou mesmo de uma *res cogitans* seria uma metáfora inadequada – uma metáfora cuja inadequação teria sido responsável por

quase todos os paradoxos com que se defrontaram as filosofias pós-cartesianas ao refletir sobre a distinção entre o físico e o mental. O termo "substância imaterial" parece remeter-nos, de imediato, a uma contradição. Com efeito, ele nos remete a um tipo de metáfora onde o visível e o extenso parecem participar: a idéia de substância parece estar, inequivocadamente, ligada à idéia de extensão. Se rejeitamos esta metáfora, verificamos que nem mesmo as assimetrias entre o físico e o mental insistentemente apontadas por Descartes seriam suficientes para se instituir uma separação entre dois *tipos de substâncias*. O Cartesianismo estaria defendendo uma teoria da *superveniência* do mental sobre o físico, uma teoria que impede que se conceba o mental como uma coisa do mundo ou como uma outra *substância*. O mental *supervem* ao físico, e isto significa dizer que o físico manifestaria o mental, este último dependeria do primeiro, mas não seria redutível a ele.

Certamente a própria noção de superveniência que acabamos de introduzir oferece dificuldades conceituais, constituindo uma noção ainda bem pouco familiar. Uma maneira de torná-la mais clara, pelo menos intuitivamente, pode ser obtida se refletirmos sobre o exercício habitual de nossa própria faculdade de nos comunicarmos através da linguagem. Quando falamos emitimos sons. Estes sons certamente constituem algo físico, detectável e mensurável, mas quando falamos emitimos também significados lingüísticos. Ora, estamos dispostos a reconhecer a existência destes últimos, embora não possamos medir nem detectar significados da mesma maneira que o fazemos com coisas físicas. Há certamente uma independência categorial entre sons e significados, embora estes últimos, para poderem se manifestar, dependam da existência dos primeiros. Afirmar que o significado supervem aos sons quer dizer que ele não depende de nenhuma propriedade física especificável que poderíamos detectar nas ondas sonoras e que as dotaria de significado. Se transpusermos esta idéia para conceber as relações entre o físico e o mental chegaremos à visão de que o mental não pode ser identificado com uma propriedade específica ou com a totalidade de

propriedades que uma determinada substância física apresente. E é esta concepção que permitirá sustentar que o mental *é privado, inescrutável e incorrigível.*

A interpretação do Cartesianismo como uma espécie de precursor da teoria da superveniência do mental sobre o físico é reforçada pelas suas teses indesejáveis acerca da restrição às possibilidades e habilidades dos autômatos que embora pudessem um dia replicar todas as características físicas de um ser humano não poderiam ou não chegariam necessariamente a produzir algo parecido à vida mental autêntica[3]. Isto porque, replicar as características físicas do ser humano, seria condição *necessária* mas não *suficiente* para se produzir vida mental, ou, em outras palavras, replicar as características físicas pode não ser suficiente para que ocorra a superveniência do mental sobre o físico.

Pode a teoria da superveniência resolver o problema da interação entre o material e o imaterial? A resposta que podemos fornecer é apenas parcial. Restaria ainda a tarefa de mostrar como determinadas ações seguem-se como conseqüência da ocorrência de conteúdos mentais específicos – uma tarefa onde a filosofia da ação e a filosofia da mente teriam de se complementar. Por outro lado, é preciso assinalar que esta formulação do conceito de superveniência, na medida em que desloca o mental para o domínio do inescrutável, torna a Psicologia como ciência uma tarefa impossível: o mental não poderia ser objeto de estudo independente. A alternativa para este impasse é mostrar a possibilidade de se formular uma teoria da superveniência do mental sobre o físico que não implique uma ruptura com o materialismo - pelo menos com o materialismo entendido como proposta metodológica. Esta tarefa exige a elaboração de uma nova metáfora que permita acomodar num modelo coerente as propriedades assimétricas do físico e do mental – um trabalho que exige, como etapa preliminar, uma revisão das teorias das relações mente-corpo formuladas no decorrer do século XX e que passamos a examinar em seguida.

[3] Veja-se o *Discours de la Méthode*, 1ª parte.

II

Deixando de lado por ora as discussões de interpretação da filosofia de Descartes, é preciso assinalar que, de uma maneira ou de outra, seja através de uma má interpretação ou mesmo de uma metáfora infeliz, é a separação cartesiana que leva a formulação explícita do problema mente-corpo no decorrer da história da Filosofia. Foi o Cartesianismo que iniciou, pelo menos historicamente, uma série de desdobramentos e de especulações em torno da natureza de duas substâncias diferentes: o físico e o mental. A história de como se tentou resolver este problema percorre uma série de alternativas para se achar um possível relacionamento entre estas duas substâncias – sobretudo um relacionamento que permitisse sua comunicação causal – apesar de sua absoluta diversidade. No século XX, uma das maneiras de se tentar resolver este problema foi procurar reduzir um tipo de substância a outra, o que imprimiu à história da Filosofia o mesmo movimento pendular que encontramos nas filosofias pós-cartesianas.

Episódios mais recentes da discussão do problema mente-corpo serão encontrados em trabalhos como *A Análise da Mente*, de B. Russell e o extenso artigo de H. Feigl, "O Físico e o Mental", apenas para citar algumas obras mais relevantes. Análises mais sofisticadas, discutindo aspectos lingüísticos do problema, encontramos na obra de G. Ryle (*The Concept of Mind*) e em alguns de seus seguidores mais recentes, como por exemplo D. Dennett (*Content and Consciousness*). Abordagens defendendo o materialismo no sentido estrito encontramos na formulação clássica da chamada teoria da identidade mente-cérebro, defendida por J.J.C. Smart e U.T. Place. Por outro lado, análises defendendo o dualismo (pelo menos numa versão branda) encontramos em artigos paradigmáticos como "O Fisicalismo" de Th. Nagel.

As grandes novidades nas abordagens do problema mente-corpo no século XX parecem ter se concentrado em duas grandes vertentes. Numa primeira vertente temos as tentativas de tratar as relações

mente-corpo como um problema lingüístico ou um problema da filosofia da linguagem que, no limite, poderia ser dissolvido após uma cuidadosa revisão do vocabulário psicológico e da terminologia que usamos para referir-nos aos nossos (possíveis) eventos mentais. Esta tentativa teve como pioneiro G. Ryle (que mencionamos acima) que, sem dúvida, transpôs para suas análises, ainda que de modo tardio, as heranças do "linguistic turn" como marca dominante da Filosofia do século XX. Numa segunda vertente, encontramos as discussões acerca da possibilidade do materialismo tendo como ponto de partida a chamada Lei de Leibnitz. Na realidade as duas vertentes tendem a se encontrar numa só: as discussões da Lei de Leibnitz levam invariavelmente a paradoxos semânticos e a questão de sentenças sem-sentido, o que nos devolve a questões de filosofia da linguagem. Finalmente temos a proposta do chamado *materialismo eliminativo* que parece ter levado a uma afirmação apressada da comensurabilidade entre teorias neurológicas e teorias psicológicas. Examinaremos em primeiro lugar alguns aspectos da teoria proposta por Ryle, para em seguida relacionar a análise do vocabulário psicológico com os paradoxos semânticos originados a partir da aplicação da Lei de Leibnitz. Usaremos a última parte desta seção para discutirmos a versão do materialismo eliminativo apresentada por R. Rorty.

Ryle e a Análise Lingüística da Mente

Em *The Concept of Mind*, Ryle desenvolve uma das análises mais promissoras no sentido de esclarecer o problema das relações entre o físico e o mental. Este trabalho, publicado em 1949, inaugura uma nova era na filosofia da mente, qual seja, a da aplicação da *ordinary language philosophy* (filosofia da linguagem comum) como metodologia para tentar resolver os problemas que envolvem a natureza do mental. Nele, Ryle argumenta que as investigações empreendidas até então assentam-se sobre um erro colossal que deu origem a uma série interminável de equívocos.

O problema central que serve de ponto de partida para a filosofia da mente, qual seja, o problema das relações mente-corpo não constitui nada além de um pseudoproblema que resultaria das confusões e abusos oriundos da linguagem que utilizamos para nos referir a fenômenos mentais. Uma análise cuidadosa da maneira pela qual falamos acerca da mente levaria a dissipar todas essas confusões e a eliminação de todas as disputas metafísicas que aí estariam pressupostas, recolocando-nos numa atitude preliminar ao dualismo, materialismo e os outros "ismos" nos quais se expressam rivalidades metafísicas. *Teorias* acerca da natureza do mental nada mais seriam do que criações daqueles que teriam, inadvertidamente, deixado se enredar pela confusão lingüística pressuposta no vocabulário psicológico cotidiano.

A estratégia de Ryle baseou-se na análise lingüística e semântica, que consistia em tentar evidenciar as perplexidades e paradoxos que surgiam da maneira como falamos acerca de mentes, sensações e vários tipos de pensamentos: seu pressuposto era o de que a análise lingüística faria com que se dissipassem *todos* os problemas da filosofia da mente – o que certamente revelou-se uma posição insustentável. O pressuposto tácito desta estratégia era uma forma branda de "behaviorismo lógico", ou seja, a visão de que a atribuição de estados mentais a um organismo deve depender, em última instância, da observação de seu comportamento manifesto.

Nesta perspectiva, o problema mente-corpo nada mais seria do que um equívoco: o físico e o mental constituiriam diferentes categorias lógicas e, neste sentido, seria inútil tentar identificá-los ou encontrar algum tipo de conexão entre estas duas esferas. O eventual insucesso da teoria da identidade mente-cérebro estaria assim perfeitamente justificado bem como estariam afastadas as preocupações que poderiam decorrer da necessidade de postular uma ontologia específica para o mental.

Mas será possível estabelecer uma distinção lógica ou conceitual entre termos físicos e termos mentais? E seria ela suficiente para sustentar

que teorias da identidade mente-cérebro devem ser abandonadas? Será que todas as conexões entre o físico e o mental – mesmo que concebidas apenas como produto de nosso vocabulário psicológico – levam necessariamente a *transgressões categoriais*?

Um bom ponto de partida para esta discussão seria apontar a existência de alguns casos anômalos que sugerem que a distinção entre o físico e o mental nem sempre pode ser realizada de maneira nítida. Tomemos por exemplo as noções de distância e de *medida* de distâncias, tais como, "quilômetros" ou "graus centígrados". Como situar a idéia de "quilômetro", seja no vocabulário físico ou no vocabulário mental? Podem os quilômetros que existem entre a Terra e a Lua serem identificados com algo *físico* do mundo? Certamente que não. Mas, por outro lado, não é possível conceber a *distância* a não ser como algo físico. Quilômetros e graus centígrados teriam uma existência tênue entre o físico e o mental – uma existência que, paradoxalmente, parece mais tênue do que aquela dos pensamentos, crenças e desejos. Certamente não podemos borrifar tinta num pensamento, numa crença ou num desejo, mas podemos dizer, por exemplo, que um pensamento ocorre na minha cabeça. Explosões (que são coisas físicas) também podem ocorrer em algum lugar do espaço. Uma dor pode ser intensa – da mesma maneira que afirmamos que uma chama de fogão é intensa ou tem intensidade. Um desejo pode embrulhar o meu estômago, o mesmo que ocorre quando como um sanduiche de alho. Entretanto, a análise lingüística dificilmente poderia classificar como transgressões categoriais as afirmações de que o pensamento ocorre, a dor é intensa e um desejo embrulha meu estômago: estas sentenças não só fazem sentido como o fazem por transitar entre o físico e o mental. O resultado é inverso no caso dos quilômetros e dos graus centígrados onde a análise lingüística reverte nossa tendência habitual de situá-los do lado do vocabulário físico.

Um outro exemplo interessante é a análise do termo "voz". Devemos situar vozes no vocabulário físico ou no vocabulário mental?

Quando afirmo "ouço uma voz" ou "perdi minha voz" ou mesmo "ele ouve vozes", será que posso situar "voz" em domínios distintos? Que sentido tem tratar "voz" como coisa física quando afirmo "perdi minha voz"? Terá sentido a afirmação "perder a voz" entendida como perder *um objeto físico*? Por outro lado, a idéia de voz como coisa física não pode ser abandonada quando gravo minha voz numa fita magnética e a mando para um amigo em Londres. Será então a voz que eu perco uma metáfora que significa, na realidade, a perda temporária de uma disposição? Neste caso, para resolver nossa dificuldade teríamos apenas que recensear os diferentes sentidos da palavra "voz": haveria uma alternância entre o sentido físico e o sentido mental. Mas isto rapidamente leva a paradoxos, pois, se assim fosse, a voz que eu emito não poderia ser a voz que eu perdi ontem; uma seria física e outra seria mental. Talvez o mesmo se aplicasse às vozes que ouço quando tenho um surto de esquizofrenia: não *penso* vozes, *ouço* vozes, e como poderia saber se elas são físicas ou mentais?

As dificuldades que relatamos parecem forçar a seguinte conclusão (pelo menos temporariamente): ou bem devemos abandonar o programa teórico ryleano de separar o vocabulário físico do vocabulário mental e com ele o seu pressuposto fundamental de que é sempre possível detectar e eliminar as transgressões categoriais ou bem admitir que a passagem do físico para o mental, expresso em sentenças híbridas, não significa necessariamente o uso indevido da linguagem e, conseqüentemente a geração de paradoxos semânticos e transgressões categoriais. Assim sendo, não parece necessário, pelo menos por enquanto, abandonar as teorias da identidade mente-cérebro com base na premissa de que elas seriam forçosamente condenadas ao fracasso simplesmente pelo fato de elas incorporarem e explicitarem transgressões categoriais e passagens entre o físico e o mental.

A Teoria da Identidade e a Lei de Leibnitz

Não é difícil esboçar uma continuidade entre a análise da terminologia psicológica empreendida por Ryle com as questões que envolvem a aplicação da Lei de Leibnitz na discussão das relações mente-cérebro. O caminho pode ser traçado se considerarmos as conseqüências da aplicação desta lei na comparação entre o físico e o mental. Esta comparação leva quase que inevitavelmente a proposições sem sentido: estaríamos aqui diante de paradoxos semânticos que poderiam facilmente ser relacionados com a idéia de transgressão categorial. Voltaremos a tocar neste assunto no final desta seção.

A Lei de Leibnitz foi freqüentemente invocada como uma séria objeção às teorias fisicalistas e às teorias da identidade mente-cérebro: ela diz respeito ao problema de relacionar as duas "substâncias", a física e a mental. Os defensores de versões recentes da identidade mente-cérebro[4], sustentam a existência de uma identidade contingente entre estados mentais e estados cerebrais. Isto significa dizer que esta seria uma identidade especial, ocasional e não necessária, como a identidade lógica que faz com que a proposição $2 + 2 = 4$ seja válida em todos os mundos possíveis. Ora, a Lei de Leibnitz diz que se duas coisas ou entidades são idênticas, a elas deve-se poder atribuir o mesmo conjunto de predicados. É neste momento que encontramos uma série de paradoxos semânticos na medida em que desejamos atribuir predicados idênticos a estados mentais e a estados cerebrais. Pois se estes últimos ocorrem no cérebro, ou são resultados da atividade do cérebro, a eles é legítimo atribuir propriedades características dos neurônios como por exemplo "umidade" ou "capacidade de transmitir correntes elétricas", "ocorrer a 5cm de distância do hemisfério esquerdo da massa encefálica" e assim por diante – propriedades que não faria sentido atribuir a estados mentais. A identidade mente-cérebro estaria gerando, inevitavelmente, um conjunto de paradoxos semânticos. Ora, não serão estes paradoxos idênticos à situação de transgressão categorial de que falávamos

[4] A referência é a autores como U.T. Place e J.J.C. Smart.

há pouco? E se isto é verdade, não seria legítimo supor que a própria tese da identidade mente-corpo deve ser abandonada, uma vez que ela resultaria do uso indevido da terminologia psicológica?

Embora à primeira vista um argumento baseado na Lei de Leibnitz possa nos induzir a uma defesa do dualismo, ele não é tão seguro assim: com efeito, eu poderia perfeitamente dizer, por exemplo, que a cor de minhas meias é azul, mas certamente não estaria autorizado a dizer das coisas azuis que elas "têm a cor das minhas meias". Em outras palavras, o que estou dizendo é que o predicado "ser azul" não é necessariamente sinônimo do predicado "ter a cor das minhas meias" embora o que de *fato* esteja sendo atribuído a um determinado objeto ou entidade coincida. Assim sendo, é perfeitamente possível conceber que estados mentais e estados cerebrais partilhem das mesmas propriedades embora os predicados que sejam atribuídos a uns e a outros difiram, ou melhor, não sejam sinônimos – uma situação que poria em questão a aplicabilidade da Lei de Leibnitz como critério para sustentar a dessemelhança entre o físico e o mental.

Lamentavelmente, esta possível refutação da Lei de Leibnitz não nos coloca numa posição inteiramente confortável, ou seja, ela nos parece insuficiente para sustentarmos que encontramos uma refutação definitiva do dualismo. Os filósofos dualistas insistem na separação entre dois tipos de substância e são muito hábeis em argumentar a favor deste tipo de partição ontológica. Um forte argumento a favor da dualidade das substâncias baseia-se na questão da identidade pessoal. Se à proposição "Eu sou o JFT" correspondesse um objeto ou um estado cerebral ela deixaria de ter sentido e a identidade pessoal torna-se-ia algo impossível, pois isto significaria dizer que este estado de coisas ou este estado cerebral poderia, na sua contingência, ocorrer a algum outro sujeito empírico (ou cérebro) e, embora eu pudesse admitir que esse outro sujeito poderia ter a mesma perspectiva de mundo que eu tenho, não faria sentido supor que ele *também* é o JFT. Haveria assim, pelo menos um estado mental ao qual não poderia corresponder um estado

cerebral, o que já implicaria num primeiro abalo à teoria da identidade mente-cérebro. Mas a este argumento, extremamente convincente à primeira vista, poderíamos objetar o seguinte: se a *res cogitans* é por essência indivisível, como pode ela "partir-se" ou "dividir-se" em várias partes e assim constituir indivíduos com identidades únicas e irreprodutíveis?[5] Se o físico e o mental são inteiramente separados, como posso saber que este é *meu* corpo e não o corpo de outrem? O grande problema do dualismo parece residir no fato de ele não nos dizer nada acerca da natureza do mental, além do fato de ele *ser distinto* do físico. Para o filósofo dualista trata-se, antes de mais nada, de postular uma separação e esta faz com que a Filosofia acabe onde ela deveria estar começando.

Rorty e o Materialismo Eliminativo

E a contrapartida do dualismo, qual seja, o materialismo? Este também nos oferece uma série de dificuldades conceituais. Se sua versão ortodoxa, expressa na teoria da identidade mente-cérebro dificilmente pode ser sustentada, talvez valha a pena examinar algumas de suas variantes mais recentes. Tal é, por exemplo, o materialismo eliminativo que sustenta que a única realidade do mental é a sua base neurofisiológica e que, se ainda utilizamos termos intencionais ou psicológicos para nos referirmos à nossa própria atividade mental, estes constituem apenas uma terminologia provisória, que progressivamente será substituída pelo vocabulário da Neurologia à medida em que esta for estabelecendo ou mapeando as relações entre o cérebro e os estados mentais. No limite, o psicológico será inteiramente substituído pelo neurofisiológico quando a ciência do cérebro estiver concluída – mesmo que num futuro distante. A grande diferença entre o materialismo eli-

[5]Este problema parece ter preocupado vários teóricos da identidade pessoal. Veja-se por exemplo, Perry, J. (1972). Mesmo supondo que cada um de nós tenhamos um pouco da *res cogitans* e que ela permaneça indivisível em cada um de nós, é preciso notar que aqui já encontramos uma implícita referência à espacialidade e à extensão.

minativo e as teorias da identidade tradicionais consiste em afirmar a existência e a possibilidade desta evolução da Psicologia em direção à Neurologia – uma evolução parecida àquela que estabelece a passagem do estágio metafísico para o estágio da ciência positiva que encontramos na formulação ortodoxa da doutrina positivista do progresso do conhecimento da humanidade.

O risco em que incorre esta posição constitui-se na sua recusa em querer explicar a natureza do mental: a explicação transforma-se na negação de qualquer ontologia própria que poderia ser atribuída aos estados internos e à terminologia psicológica que utilizamos para descrevê-los. Mas esta posição suscita ainda outras objeções imediatas. Em primeiro lugar é preciso assinalar que o *materialismo eliminativo* constitui apenas um imenso *programa teórico*: até agora nenhuma substituição ou mapeamento de termos psicológicos em termos neurológicos foi alcançado e mesmo que o fosse restaria saber se ele seria aceitável... Por outro lado, eliminar o mental para explicá-lo coloca o risco de uma transgressão categorial: seria o mesmo que formular uma explicação histórica para a morte de Joana D'Arc utilizando-se de leis da física que determinam como e porque ocorre a combustão da lenha empilhada numa fogueira...

A este respeito, vale a pena comentar a posição de R. Rorty no seu livro mais influente, *Filosofia e o Espelho da Natureza*. Neste trabalho Rorty insiste na tese de que a idéia de "mente" entendida como algo distinto e constituindo uma esfera própria de investigação não passa de uma invenção dos filósofos do século XVII. Sua crítica do estatuto próprio do mental aproxima-o de filósofos como Ryle e Wittgenstein, que tentaram exorcizar o chamado "fantasma na máquina" (the ghost in the machine). Esta crítica corre paralelamente aos seus ataques às epistemologias fundacionalistas que concebem que a "mente é um grande espelho, contendo várias representações, algumas corretas, outras não" (Rorty, R. 1979, p. 12). A tarefa da epistemologia seria obter representações corretas, estabelecendo assim a teoria do conhe-

cimento. Esta concepção teria sido igualmente herdada pela filosofia analítica contemporânea e até mesmo pelas tentativas de estabelecer teorias *empíricas* do conhecimento tais como a epistemologia genética de Piaget ou mesmo as teorias da ciência cognitiva contemporânea.

Mas o que causa estranheza no livro de Rorty são seus capítulos iniciais. Neles, ele nos convida a imaginar uma comunidade que não usaria termos intencionais para descrever suas próprias atividades mentais. Com isto ele tenta livrar-se dos eventuais problemas envolvidos no mapeamento entre o mental e o neurológico – uma conseqüência da epistemologia representacionalista que teria levado à equivocada invenção da mente. Ora, ocorre que isto não parece resolver os problemas, uma vez que com esta manobra Rorty acaba introduzindo, subrepticiamente, precisamente os pressupostos da epistemologia representacionalista que ele tanto visa refutar: afinal, não estará ele pressupondo que o vocabulário neurológico é a única e correta representação do mundo e com isto reintroduzindo o "espelho da natureza"? Ou teria ele esquecido que mesmo uma descrição neurológica da "atividade mental" constitui por si só uma *representação*? E por que privilegiar esta forma de representação se queremos desconstruir o "espelho da natureza"? O desejo de delimitar o lugar cósmico privilegiado de onde poderíamos ter a descrição final da natureza do mental estaria reaparecendo.

O rápido balanço da situação deste debate metafísico e alguns de seus desdobramentos na filosofia da mente contemporânea não parece nada animador. Conforme notamos, as duas alternativas que apresentamos rapidamente acima conduzem-nos a dilemas teóricos. Por outro lado, o materialismo eliminativo, entendido como uma versão modificada da teoria da identidade, parece não ter contribuído para a formulação de uma teoria coerente das relações mente-cérebro. Isto é o que nos revela a crítica que esboçamos a um de seus maiores defensores – por mais oblíqua que esta crítica tenha sido.

Mas o que torna o materialismo uma posição particularmente difícil de ser sustentada? Se queremos estipular uma teoria materialista da

mente de acordo com bases científicas é inevitável que estados mentais devam se conformar às leis da Física. Se estas últimas estão corretas e, se, além disto, estados mentais ocorrem no tempo (o que parece difícil ou quase impossível de negar) então devemos igualmente supor que estados mentais ocorrem *no espaço*, ou seja, que eles devam ter algum tipo de localização espacial.

Mas, como notamos acima, a não ser que a Lei de Leibnitz tenha sido refutada – o que não parece ter sido o caso até o momento – facilmente enveredamos por paradoxos semânticos ao tentar localizar estados mentais no espaço: que sentido haveria em afirmar que meu sonho ocorreu a 5cm do hemisfério esquerdo do meu cérebro? Ou que a minha ansiedade está localizada a 5cm do hemisfério esquerdo do meu cérebro?

Ora, uma possível saída para esta dificuldade pode ser vislumbrada se correlacionarmos a aplicação da Lei de Leibnitz com a idéia de transgressão categorial – uma posição que viemos sugerindo no decorrer desta seção. A aplicação da lei de Leibnitz e a conseqüente geração de paradoxos semânticos ilustraria um caso-limite (e por isso mesmo surpreendente) de transgressão categorial: a origem dos paradoxos semânticos residiria na forçada aproximação entre a linguagem do mental e a linguagem do físico resultante da atribuição do mesmo conjunto de predicados a entidades que supostamente são idênticas.

Mas, se a aplicação da Lei de Leibnitz não nos conduz a nada além de um caso-limite de transgressão categorial, então ainda há alguma esperança para se sustentar uma posição materialista: conforme vimos, a análise da linguagem e a tentativa de separação entre as esferas do físico e do mental através da eliminação das transgressões categoriais não nos permite o traçado de uma linha divisória precisa entre estes dois domínios. Esta quase impossibilidade revela que a própria noção de transgressão categorial e sua aplicabilidade nas relações entre o físico e o mental talvez não seja tão segura quanto se desejaria supor.

No caso específico da questão da localização dos estados mentais, a Lei de Leibnitz (entendida como caso limite de geração de transgressões categoriais) só gera paradoxos semânticos em contextos particularmente limitados: com efeito, não faz sentido afirmar que o meu sonho ocorreu a 5cm do hemisfério esquerdo do meu cérebro , ou no neurônio que convencionei ter o número 235. Contudo, não parece ser paradoxal afirmar que o meu sonho ocorreu no quarto onde eu dormia, nem tampouco afirmar que o meu sonho ocorreu *no mundo*. A idéia de transgressão categorial rapidamente se dissolve, e com ela os paradoxos da Lei de Leibnitz. Se meu sonho não tivesse ocorrido no mundo, como poderia falar dele quando retorno à vida desperta? A questão da possibilidade de se falar da localização de estados mentais é fundamental para se sustentar o materialismo – pelo menos o materialismo enquanto proposta metodológica – e sobre este assunto voltaremos a falar na seção 4 quando introduzirmos os modelos conexionistas como metáforas privilegiadas para se conceber as relações entre o físico e o mental.

III

A comparação entre posições radicalmente opostas como o materialismo e o dualismo leva-nos de volta à nossa proposta inicial: o problema mente-corpo – na qualidade de problema metafísico – talvez não possa receber uma solução definitiva, restando-nos apenas a possibilidade de conceber uma metáfora adequada através da qual algumas das dificuldades conceituais presentes nestas diversas posições possam ser parcialmente removidas. E é precisamente esta a proposta que se origina da Inteligência Artificial.

Em 1975 H. Putnam publica o artigo "Minds and Machines" visando reavaliar o problema das relações entre o físico e o mental. Neste artigo Putnam desenvolve duas teses que correm paralelamente. A primeira consiste numa defesa do funcionalismo *a la máquina de Turing* que visa estabelecer um psicoparalelismo sem lançar mão de hipóteses

metafísicas. A segunda consiste em defender uma possível ligação entre o funcionalismo e o materialismo, uma posição que será batizada de *identidade teorética* entre estados mentais e estados cerebrais.

O funcionalismo, enquanto tese geral defendida pelos teóricos da Inteligência Artificial, sustenta que estados mentais são definidos e caracterizados pelo *papel funcional* que eles ocupam no caminho entre o *input* e o *output* de um organismo ou sistema. Este papel funcional caracteriza-se seja pela interação de um estado mental com outros que estejam presentes no organismo ou sistema, seja pela interação com a produção de determinados comportamentos. O funcionalismo consiste, assim, num nível de descrição onde é possível *abster-se* ou *suspender-se* considerações acerca da natureza última do mental. É também com base nesta tese de que estados mentais definem-se pelo seu papel funcional que um sistema pode apresentar predicados mentais independentemente do tipo de substrato físico do qual eles poderiam eventualmente resultar. Um mesmo papel funcional que caracteriza um determinado estado mental pode se instanciar em criaturas com sistema nervosos completamente diferentes, e nesse caso diremos que eles estão no *mesmo* estado mental.

Ora, o funcionalismo não implica no materialismo, mas também não é incompatível com este último. E é esta possibilidade que Putnam explora no seu artigo. Sua percepção é que esta compatibilidade pode se tornar possível após uma análise lingüística do problema das relações mente-cérebro. Não se trata de mostrar que o materialismo consiste na solução para o problema mente-cérebro, mas de mostrar que ele é *possível* uma vez que tenhamos optado pelo funcionalismo como ponto de partida – um ponto de partida que nos possibilita, desde o início, sustentar um psicoparalelismo sem introduzir hipóteses metafísicas *ad hoc*, como um *deus ex machina* ou uma harmonia pré-estabelecida.

Comecemos pela análise deste psicoparalelismo. O ponto de partida do psicoparalelismo é dado pelo computador: as relações entre o físico e o mental (ou entre o cérebro e a mente) podem ser concebidas como

uma relação entre o *software* e o *hardware* de uma máquina. Putnam concebe o funcionamento mental ao modo de uma máquina de Turing – afinal, todos os computadores digitais são essencialmente máquinas de Turing.

A máquina de Turing é fundamentalmente um processador de símbolos: uma máquina virtual que pode ser representada em termos de uma longa fita de papel ou de qualquer outro material que contenha símbolos e marcas a intervalos regulares, formando pequenos quadrados[6]. Imaginemos agora que podemos estipular uma espécie de marcador ou um ponto fixo em relação ao qual pudéssemos mover a fita de papel para a esquerda ou para a direita, e suponhamos, igualmente, que o nosso marcador tenha também um dispositivo que permita reconhecer se num determinado quadrado há um símbolo ou não. O marcador pode também imprimir e apagar símbolos que aparecem na fita e movê-la para a esquerda ou para a direita, dependendo do símbolo que aparece impresso num determinado quadrado.

Na fita que imaginamos podemos convencionar símbolos de dois tipos: símbolos escritos em letras minúsculas e letras maiúsculas. Mover a fita para a esquerda ou para a direita (e num número determinado de quadrados) dependerá do símbolo em maiúsculas que seja identificado pelo marcador. Além de mover a fita em determinadas direções, o símbolo em maiúsculas pode significar que o marcador deve imprimir ou apagar um símbolo num certo quadrado. Com este tipo de máquina virtual pode-se, rigorosamente falando, executar qualquer tipo de operação simbólica – e é precisamente esta a inovação introduzida por Turing, que com a invenção de sua máquina, forneceu uma espécie de princípio geral para a construção de qualquer tipo de computador.

Ora, a idéia de Putnam é que a máquina de Turing fornece-nos uma excelente analogia ou um bom modelo para concebermos a relação mente-cérebro: de um lado, há um conjunto de regras abstratas (ins-

[6]Para uma explicação mais detalhada da noção de máquina de Turing, veja-se Teixeira, J. (1990), cap. 2.

truções) e de outro, a realização física dessas regras obtidas pelos diferentes estados da máquina. Assim, a analogia consiste basicamente em estabelecer uma correlação entre estados mentais (pensamentos) e o *software* (conjunto de instruções da máquina ou o programa do computador) de um lado e entre estados cerebrais e o *hardware* ou os diferentes estados físicos pelos quais passa a máquina ao obedecer as instruções. O psicoparalelismo torna-se assim concebível com base neste esquema conceitual – um psicoparalelismo que dispensaria qualquer tipo de pressuposição metafísica que seria responsável pela possibilidade de interação entre o físico e o mental.

O esquema de interação entre o físico e o mental na analogia *software/hardware* proposta por Putnam é particularmente sugestiva mas não imune a críticas. Em primeiro lugar, a analogia de Putnam não escapa às objeções comuns que se tem levantado contra os modelos funcionalistas em geral. Ao definir a máquina de Turing como máquina virtual cujo substrato material pode ser de qualquer natureza e ao definir a atividade mental como dependente do desempenho de um determinado conjunto de funções, Putnam incorre na objeção mais freqüente que se levanta contra o funcionalismo em geral: o argumento da China. Se cada habitante da China (que tem mais de um bilhão deles) desempenhar o papel que normalmente seria atribuído a um neurônio, estaríamos em condições de atribuir à China atividade mental – o que certamente seria contra-intuitivo. (Exceto talvez para aqueles humanistas que gostariam de definir algo parecido com o "espírito de um povo"; mas isto certamente não passaria de uma caricatura).

Em segundo lugar – e esta parece ser a objeção mais séria ao modelo de Putnam e aos modelos funcionalistas em geral – há a chamada questão da individuação dos estados mentais. Como se estabeleceria num modelo funcionalista *a la* máquina de Turing a identidade específica de um determinado estado mental? Não ficaria ela por conta da interpretação que um observador externo à máquina poderia atribuir aos símbolos que são manipulados por esta última? Com efeito,

se sustentamos a independência entre estados físicos e estados computacionais ou estados do *software*, podemos imaginar organismos com exatamente o mesmo programa e com o mesmo tipo de estados mentais (ou computacionais), apesar deles diferirem no que diz respeito à sua fisiologia e à sua possível interação com o meio ambiente. Até aqui nada mais estaríamos fazendo do que enunciar a tese central do funcionalismo. O problema surge na medida em que podemos imaginar duas criaturas com estados funcionais de tipo idêntico mas diferindo no que diz respeito ao *conteúdo* de seus pensamentos. No caso de uma máquina de Turing é possível conceber que um mesmo programa simule duas situações completamente diferentes mas que se apresentarão isomórficas do ponto de vista computacional. É difícil de imaginar, mas não impossível de ocorrer, que se possa utilizar um mesmo programa computacional para simular a guerra do Iraque contra o Kuwait ou um jogo de xadrez. Do ponto de vista funcional estes programas seriam indistinguíveis, mas no que diz respeito aos conteúdos mentais que eles estariam supostamente simulando, seriam totalmente diferentes. Neste caso, chegamos a uma estranha idéia de um psicoparalelismo: um mesmo conjunto de estados de *software* e de *hardware* corresponderia a conjuntos inteiramente distintos de conteúdos mentais. O problema é que o modelo da máquina de Turing é excessivamente geral, não permitindo a individuação de estados mentais a não ser por uma atividade externa de atribuição de conteúdos mentais aos estados de *software* e de *hardware*.

Para se ter uma idéia mais prática do problema de que estamos tratando basta imaginar que temos um computador rodando um determinado programa e que a uma certa altura resolvamos reduzir o programa a uma linguagem *Assembler*. Em seguida, executamos mais um passo e reduzimos a linguagem *Assembler* à linguagem de máquina. Se alguém quiser reconstruir o programa que estava sendo rodado a partir da linguagem de máquina encontrará grandes dificuldades para fazê-lo, na medida em que poderá haver uma ou mais linguagens de

alto nível compatíveis com a mesma linguagem de máquina. Mas a dificuldade poderá ser ainda agravada se imaginarmos dois computadores rodando linguagens diferentes e simulando duas situações incomparáveis exibindo a mesma linguagem de máquina após um processo de compilação. Esta é sem dúvida uma situação-limite e muito pouco provável, mas não impossível do ponto de vista lógico, o que põe a perder a possibilidade de se estabelecer um paralelismo psicofisiológico com base no modelo da máquina de Turing como é pretendido por Putnam.

Examinemos agora, por uma questão de completude, a tese da *identidade teorética* entre estados mentais e estados cerebrais. Conforme dissemos acima, com esta tese Putnam pretende mostrar que o materialismo é pelo menos um horizonte *possível* – uma teoria que não tem de necessariamente forçar-nos a cair em contradições ou paradoxos insolúveis. Sua reflexão parte do significado da identidade, e sua formulação aproxima-se de uma variação da teoria da identidade contigente. Só que desta vez trata-se de sustentar que o enunciado 'estados mentais = estados cerebrais' pode vir a se tornar um enunciado inteligível no interior de uma teoria futura acerca do funcionamento mental e cerebral da mesma maneira que o enunciado água = H_2O. Este último enunciado, ou seja, esta relação de identidade, não faria sentido antes da descoberta da eletrólise e da teoria que a acompanha. O mesmo poderíamos supor do enunciado 'estados mentais = estados cerebrais'.

Contrariamente à afirmação de que o enunciado 'estados mentais = estados cerebrais' estabelece apenas uma *correlação* e não uma *identidade* por tratar-se de um enunciado sintético e não analítico, Putnam argumenta que não se pode estabelecer uma distinção absoluta entre enunciados analíticos e sintéticos. Com efeito, esta distinção tem sido esmaecida ao longo da história à medida em que se verificou que certos enunciados supostamente analíticos não o são.

Da mesma maneira, alguns enunciados que supostamente não tinham sentido ou eram semanticamente desviantes passaram a tê-lo ao longo do processo histórico. Por exemplo as sentenças[7]:

[7]Estes exemplos foram tirados do artigo de Putnam (1975).

a) Estou a milhares de milhas de distância de você

ou,

b) Ele está na metade do seu sonho.

No caso de a) a sentença certamente não faria sentido na Grécia antiga onde a idéia de "milhares de milhas de distância" seria ininteligível dadas as condições de mensuração de distâncias e os meios de locomoção existentes. No caso de b) a sentença seria ininteligível até que se tivesse inventado o eletroencefalograma. Assim sendo, argumenta Putnam, é perfeitamente plausível que enunciados do tipo:

c) O estado mental "ψ" é idêntico ao estado cerebral "ϕ".

torne-se um enunciado inteligível e não apenas uma correlação, da mesma maneira que enunciados como:

d) Luz é radiação eletromagnética

ou,

e) Água é H_2O.

O que torna esses enunciados inteligíveis, e não apenas correlações, é o fato de eles ocorrerem no interior de uma teoria científica – uma teoria mais ampla que torna a identificação teórica uma possibilidade real.

Ora, que podemos dizer da idéia de identificação teorética? A idéia parece atraente à primeira vista mas não é imune a objeções. Para começar, a idéia de "estar a milhares de milhas de distância" pode ser uma idéia difícil de ser imaginada (diante de meios de locomoção ainda precários) mas isto não quer dizer que ela não possa ser *concebida* – o que torna o enunciado perfeitamente inteligível em qualquer época. Assim sendo, o enunciado (a) acaba se tornando um contra-exemplo ao ponto de vista de Putnam, ou, no mínimo, um exemplo infeliz.

Mas não são apenas estes detalhes que nos interessam aqui. A própria idéia de identificação teorética pode nos levar rapidamente a objeções. Se a identificação entre estados mentais e estados cerebrais depende de uma aposta no desenvolvimento futuro da ciência, o que me garante que as teorias científicas irão de fato contribuir para tornar este enunciado inteligível e não vice-versa?

Putnam argumenta no sentido de mostrar que a distinção entre analítico e sintético é na realidade algo tênue – a história tem mostrado que esta linha divisória não é nítida. Mas aqui parece ter havido sempre uma direção: o que era analítico mostrou-se, em última análise, sintético. O conjunto de proposições analíticas tornou-se cada vez menor. Ora, a direção inversa não parece ser provável: não se tem notícia de que uma proposição sintética tenha se revelado analítica. Se enunciados do tipo 'água = H_2O' ou 'luz = ondas eletromagnéticas' tornaram-se analíticos no interior de teorias científicas, permitindo a intersubstitutividade *salva veritate* de termos em enunciados destas teorias é preciso não esquecer que isto se tornou possível após verificação empírica – é esta verificação que dá a estes enunciados o caráter de *definições* ou enunciados aparentemente analíticos. No caso do enunciado 'estados mentais = estados cerebrais' é bem pouco provável que esta verificação possa ser feita ou mesmo aceita na qualidade de uma verificação que nos permita assumir este enunciado como uma definição ou ponto de partida para uma teoria científica. Assim sendo, a identidade teorética pretendida por Putnam torna-se, na realidade, uma fantasia teórica: uma fantasia que toma como ponto de partida precisamente aquilo que a teoria científica precisaria resolver para que este ponto de partida pudesse ser plenamente aceito.

IV

Resta-nos agora examinar um outro modelo – ou uma outra metáfora – fornecida pela Inteligência Artificial e verificar em que sentido esta pode contribuir para esclarecer os problemas envolvidos na relação

mente-corpo: trata-se do funcionalismo neurocomputacional ou conexionismo, iniciado com os trabalhos de von Neumann, McCulloch & Pitts, Hinton e, especialmente Rumelhart e McClelland.

O funcionalismo neurocomputacional não endossa a visão de que processos mentais possam ser estudados como computações abstratas, independentemente de sua base física e do meio ambiente onde se situa o organismo ou o sistema onde elas ocorrem. Conhecimentos acerca do funcionamento do cérebro e conhecimentos sobre computação devem convergir no estudo da natureza dos estados mentais. O cérebro humano é visto como um dispositivo computacional em paralelo que opera com milhões de unidades computacionais chamadas "neurônios" ou "neuron-like units". Computadores e cérebros são sistemas cuja função principal é processar informações e assim pode-se utilizar redes artificialmente construídas para simular esse processamento. Tais redes constituem um intrincado conjunto de conexões entre os "neurônios" ou "neuron-like units" que estão dispostos em camadas hierarquicamente organizadas. Dado um determinado *input*, diferentes estados mentais podem ocorrer como conseqüência de mudanças nas conexões, que podem ser inibidas ou ativadas, variando de acordo com a interação do sistema com o meio ambiente e com seus outros estados internos. As conexões entre unidades estimuladas via *inputs* externos geram os chamados *padrões de conectividade*. Padrões de conectividade estocam informação acerca do que um sistema "sabe" num determinado momento.

De acordo com este modelo, a formação e modificação de padrões de conectividade ocorrem em função da experiência. Modificações na maneira de representar conhecimento ou "gerar estados mentais" podem ocorrer seja pelo desenvolvimento de novas conexões num determinado padrão de atividade do sistema, seja pela extinção de algumas conexões já existentes ou até mesmo pela modificação de pesos e valores de ativação/inibição em conexões já existentes.

Contudo, é preciso assinalar que máquinas conexionistas não constituem apenas grandes processadores em paralelo. O paralelismo não é condição suficiente para definir uma máquina conexionista. A própria noção de computação envolvida nestes sistemas é bastante diferente daquela que encontramos nos computadores baseados no processamento serial de informação e na máquina de Turing. Podemos afirmar que sistemas conexionistas baseiam-se num outro tipo de máquina virtual, a máquina de Boltzmann, inspirada num modelo termodinâmico.

A máquina de Boltzmann é composta de uma série de unidades simples operando em paralelo e conectadas com unidades vizinhas através de ligações bidirecionais. Estas ligações recebem um determinado peso que pode ser positivo ou negativo. Suponhamos agora que a um determinado momento concebamos cada uma das unidades como representando informações recebidas através de um determinado *input*. Uma determinada unidade é então ativada na medida em que ela "acredita" que aquela informação seja verdadeira. Duas unidades que representam informações contraditórias serão ligadas por uma conexão de peso negativo, enquanto que unidades que representam hipóteses coincidentes tenderão a incrementar o peso de sua conexão. Em outras palavras, as ligações permitem que as unidades individuais se excitem e se inibam entre si de uma maneira sistemática. O estado de uma unidade num determinado momento dependerá, em parte, do estado de todas as outras unidades com a qual ela está ligada. E essas unidades, por sua vez, serão influenciadas ainda por outras unidades com as quais elas estão conectadas no interior da rede. A produção de um determinado *output* dependerá assim de um processo interativo de ajustamento mútuo de inibições e excitações até que uma decisão final seja atingida – a decisão que chamamos de "decisão comunitária". Este processo de ajustamento é também chamado de "processo de relaxamento", num ciclo que guarda muita semelhança com o modelo de prazer/desprazer e o princípio de constância que norteou o modelo hidráulico da mente proposto por Freud.

Mas, em que sentido pode este modelo conexionista, baseado em idéias termodinâmicas e na máquina de Boltzman, contribuir para resolver os problemas de que viemos tratando até agora? Para começar, é preciso assinalar que aqui encontramos uma possível resposta para o problema da individuação dos estados mentais – o problema que tornava o modelo de Putnam particularmente inadequado. Um mapeamento dos estados internos e uma correspondência com estados específicos do *hardware* pode inicialmente ser obtida se imaginarmos uma rede que pode ser treinada a formar conexões específicas de tal maneira que mudanças nas suas condições externas causem a ocorrência e a formação de padrões específicos de atividade. Esta correspondência, por sua vez, determina o conteúdo informacional específico do padrão de conectividade em questão. Encontramos aqui uma primeira diferença em relação ao modelo da máquina de Turing e a analogia proposta por Putnam: na máquina conexionista o paralelismo entre conteúdos mentais e estados do *hardware* é concebido como uma relação específica. Não há independência entre *software* e *hardware* e as características de *design* deste tipo de máquina permitem que, no limite, possamos conceber que a cada estado mental (ou de *software*) corresponda um estado cerebral (ou de *hardware*). O *design* específico destas máquinas permitem que seu *hardware* possa comportar, num espaço restrito, um número quase infinito de configurações[8].

O modelo da máquina de Turing e sua concepção de computação como operação simbólica supõe que se possa estabelecer uma autonomia entre *hardware* e *software* (ou entre estados estruturais e estados lógicos ou mentais) sem introduzir hipóteses *ad hoc* para garantir o paralelismo psicofísico. Mas, conforme vimos, o preço que se paga por esta solução é a incapacidade de resolver o problema da individuação

[8]Diremos *quase* infinito. Uma vez que as redes são físicas, o número de conexões e de estados mentais será finito. Isto explica porque podemos *conceber* o infinito mas não *imaginar* o infinito. O número de estados mentais também será finito: a tese chomskyana do número infinito de sentenças que podemos construir na linguagem não significa *ipso facto* na existência de um número infinito de estados mentais.

dos estados mentais. Este problema estaria resolvido nos modelos conexionistas, na medida em que, através deles, é possível conceber uma correspondência entre a diversidade qualitativa dos conteúdos mentais e as múltiplas configurações de *hardware* – uma multiplicidade que se torna possível pois neles há um número extraordinariamente grande de combinações a partir de conexões possíveis, sejam estas fixas ou momentâneas. A idéia de *conexão momentânea* constitui uma metáfora particularmente adequada para se conceber o fluxo e a velocidade com que se sucedem os estados mentais, além de maximizar o número de combinações possíveis. A máquina conexionista é, no limite, "como caminhar por um labirinto cujas paredes modificam sua disposição a cada passo que damos"[9].

A dependência do *software* em relação ao *hardware* significa dependência de uma arquitetura ou de um *design* específico da máquina conexionista, mesmo que este possa ser representado como a arquitetura abstrata de uma máquina virtual. Por outro lado, se este novo tipo de *hardware* vier efetivamente a ser construído, temos que considerar que a dependência de que falamos acima restringe qualitativa e quantitativamente a informação que pode ser processada.

Um exemplo deste tipo de restrição poderia ser encontrado no artigo de George Miller sobre o mágico número 7. Neste artigo Miller mostra que existe um limite constante (7) para o número de itens aleatórios[10], que podem ser relembrados quando a memória humana é solicitada. Tudo se passa como se no cérebro humano a memória fosse limitada da mesma forma que uma calculadora tem um limite de casas que ela pode comportar. A investigação de Miller ilustra o que entendemos por dependência do *hardware* ao mesmo tempo que aponta para o tipo de

[9]Esta metáfora é tirada do livros de Gleick, J. (1990), p. 21. Nesta passagem Gleick refere-se a equações não-lineares na meteorologia. Curiosamente, a passagem é seguida de uma citação de J. von Neumann.

[10]Talvez o melhor termo seria "avulsos" e não aleatórios. Os itens não podem ter nenhuma relação entre si que permita qualquer tipo de associação ou agrupamento que facilite o trabalho da memória.

pesquisa que se pode desenvolver em ciência cognitiva quando utilizamos modelos conexionistas: trata-se de investigar o tipo de relação que se estabelece entre arquitetura de *hardware* e as variedades de conteúdos mentais que podemos formar.

Neste modelo, os conteúdos mentais emergem da atividade das redes e suas conexões: encontramos aqui uma direção inversa àquela do funcionalismo tradicional onde estados mentais são atribuídos a estados do *hardware*. Ou, para empregar uma terminologia filosófica, podemos afirmar que estados mentais são supervenientes à atividade das redes. Isto marca a grande diferença entre o funcionalismo neurocomputacional e o funcionalismo tradicional: este último pressupõe a *separação* entre as substâncias extensa e inextensa, enquanto no primeiro encontramos a idéia de *separabilidade* do físico e do mental, estabelecendo assim a possibilidade de eliminar algumas das dificuldades das filosofias pós-cartesianas de que falamos na primeira seção deste artigo.

Esta concepção do mental que supervem a atividade das redes sugere que a mente nada mais é do que o resultado de um processo de auto-organização de determinados sistemas físicos com características peculiares.[11]. Isto significa que em alguns sistemas físicos o fenômeno

[11]Estes sistemas físicos serão, possivelmente, aqueles que apresentam um elevado grau ou tendência à auto-organização. Esta idéia de auto-organização é bem explorada por Gleick (1990), numa passagem onde ele nos fala das observações acerca de fenômenos no planeta Júpiter: "Um modesto mistério cósmico: a Grande Mancha Vermelha de Júpiter, um enorme oval rotativo, como uma tempestade gigantesca que nunca se move e nunca se esgota. (...) as condições de tempo extra-terrenas de Júpiter revelavam-se um dos muitos problemas que esperavam um novo sentimento das possibilidades da natureza, proporcionado pela ciência do caos. (...) um especialista em dinâmica de fluidos que via a turbulência como aleatória e ruidosa, não tinha contexto para compreender uma ilha de estabilidade em seu meio. (...) A mancha é um sistema auto-organizador, criado e regulado pelas mesmas mudanças não-lineares que criam a agitação imprevisível à sua volta. É o caos estável." (veja-se Gleick, p. 48 a 52). O exemplo mostra um sistema caótico mas ao mesmo tempo auto-regulado, alternando imprevisibilidade e estabilidade. Talvez este modelo de caos, descrito por equações não-lineares seja o melhor para entender como funcionam as redes e como elas atingem uma estabilização. Uma exploração interessante seria comparar aquilo que chamamos de racionalidade com um processo de auto-organização onde se alteram a imprevisibilidade e a estabilidade.

da superveniência pode vir a ocorrer dependendo de relações complexas de probabilidade que relacionam o sistema ao seu meio ambiente.

É também esta concepção que possibilita que indivíduos inicialmente com uma mesma rede e recebendo um mesmo *input*, possam formar conteúdos mentais diferenciados, seja do ponto de vista qualitativo, seja do ponto de vista da intensidade de uma determinada sensação: é impossível determinar *a priori* quais serão as conexões a serem ativadas e como será a distribuição dos pesos pela rede. Esta seria a contribuição dos modelos conexionistas para uma possível solução do problema dos *qualia*. Conteúdos mentais são, assim, essencialmente indetermináveis não apenas na medida em que não podemos antever quais as redes que serão ativadas por um *input* sensorial, como também pelo fato de poder se estabelecer uma diferença entre conteúdos sensoriais e conteúdos representacionais em diferentes indivíduos. Conteúdos representacionais, na medida em que supervem a atividade da rede, podem tornar-se privados e inescrutáveis[12].

Finalmente, é preciso frisar uma última vantagem do emprego de modelos conexionistas como metáfora privilegiada para esclarecer o problema das relações entre o físico e o mental: modelos conexionistas tornam possível conceber estados mentais como estados materiais sem cair nos paradoxos sugeridos pela Lei de Leibnitz. Estados mentais ocorrem no espaço, embora não possamos dizer exatamente onde eles ocorrem: eles estão em algum lugar da rede de conexões entre as unidades e na forma de um processo global do sistema. Não faz sentido afirmar "meu sonho ocorre a 5cm do hemisfério esquerdo do meu cérebro", mas faz

[12]Para se entender o que chamamos de conteúdos representacionais é preciso entender em que medida eles são distintos de conteúdos sensoriais. Esta distinção pode se tornar mais nítida no caso da percepção visual: pacientes diante de uma mesma figura de *Gestalt* ou seja diante do mesmo tipo de estimulação sensorial formarão representações distintas daquilo que estão vendo. Para uma exploração parecida desta distinção, veja-se Peacocke, C. (1983), cap. 1. Conteúdos representacionais são fortes candidatos à inescrutabilidade ou pelo menos à opacidade: não poderíamos saber, a partir de observação externa e da observação dos estímulos sensoriais quais as representações internas que o sistema ou o organismo estaria formando.

sentido afirmar que meu sonho ocorre no quarto ou meu sonho ocorre no mundo, da mesma maneira que o faz afirmar que ele ocorre em algum lugar da rede. Não podemos identificar um estado mental com um estado cerebral específico da mesma maneira que não podemos localizá-lo nem dizer que ele é o resultado de uma única e possível combinação de ativações de uma determinada rede. A identidade será sempre identidade com um determinado processo e não com um grupo específico de neurônios. Conteúdos mentais não são fenômenos localizados mas o resultado de uma arquitetura específica das redes de conexões ou de um *design* específico que instancia um determinado *software*. A produção do mental depende não de um material específico nem de uma combinação simbólica mas desse *design* específico onde a ordem semântica e a ordem causal das leis da natureza constituem um mesmo e indistinguível objeto dando lugar à representação implícita ou a um estado mental.

Representações são formadas a partir das ligações entre as *neuron-like units*, num processo de passagem do sub-simbólico ao simbólico, o que metaforicamente é comparável à passagem de uma enorme quantidade de pontos sem significado algum para a formação de uma gravura: a gravura surge da união dos pontos, é distinta destes pontos, mas é sempre decomponível num conjunto quase infinitamente grande de pontos. É este processo que se inicia ao nível microcognitivo que nos autoriza a conceber uma correlação do vocabulário neurofisiológico para o vocabulário psicológico. Esta correlação não é ainda uma tradução de um vocabulário teórico para outro, uma vez que sempre haverá um hiato entre conteúdos sensoriais e conteúdos representacionais mas representa pelo menos uma possível comensurabilidade entre estes dois tipos de teorias.

Esta é a metáfora que propomos para conceber as relações mente-corpo – uma metáfora oriunda dos modelos conexionistas da mente. É ela que permite que não precisemos romper com o materialismo enquanto proposta metodológica, acomodando uma série de dificuldades

que se originam da postulação tradicional do problema das relações entre o físico e o mental. Talvez a comensurabilidade e a tradução do vocabulário neurofisiológico para o vocabulário psicológico possa um dia tornar-se possível e cheguemos a uma situação semelhante à da Física contemporânea, onde a linguagem científica afastou-se definitivamente da descrição cotidiana dos objetos e da matéria, embora saibamos que é desse mundo comum e cotidiano que continuamos falando. Talvez o mesmo venha a ocorrer com a Psicologia: a unificação do vocabulário intencional e do vocabulário neurofisiológico apresenta-se como um horizonte distante, mas possível, quando aceitarmos que a terminologia intencional possa ser substituída por uma linguagem unificada que ainda não sabemos como poderá ser. Enquanto isto não ocorre, podemos pelo menos usufruir do conforto ontológico que o materialismo metodológico nos oferece – um conforto que nos desobriga de postular a existência de mentes como um domínio separado e que se nos apresenta tão bizarro como admitir a existência de discos voadores e seres extra-terrestres.

BIBLIOGRAFIA

Alanen, L. (1981). Descartes' Dualism and the Philosophy of Mind *Revue de Métaphysique et de Morale*, 3: 391-413.

Dennett, D. (1969). *Content and Consciousness.* London: Routledge & Kegan Paul.

Descartes, R. (1641). *Abregé des Méditations.* Ed. F. Paris, Garnier, 1967, T. I.

——. (1641). *Méditations.* Ed. F. Paris, Garnier, 1967, T. I.

——. (1637). *Discours de la Méthode.* Ed. F. Paris, Garnier, 1967, T. I.

——. (1644). *Principes de la Philosophie.* Ed. F. Paris, Garnier, 1967, T. III.

Feigl, H. (1958). The "Mental" and the "Physical", in H. Feigl, M. Scriven, & G. Maxwell (eds.). *Minnesota Studies in the Philosophy of Science*, vol. 2. Minneapolis: University of Minnesota Press.

Gleick, J. (1987). *Chaos: Making a New Science.* Tradução de W. Dutra, (1990), Rio de Janeiro Editora Campus.

Hinton, G & Anderson, J. (1981). *Parallel Models of Associative Memory.* Hillsdale, NJ. Lawrence Erlbaum.

McCulloch, W. & Pitts, W. (1943). A Logical Calculus of the ideas immanent in nervous activity *Bulletin of Mathematical Biophysics*, 5: 115-133.

Nagel, T. (1965). Physicalism. In *The Philosophical Review*, 74, 3: 339-356.

Peacocke, C. (1983). *Sense and Content.* Oxford: Clarendon Press.

Place, U.T. (1970). Is Consciousness a Brain Process? in *The Mind/Brain Identity Theory*, ed. por C.V. Borst. London: The Macmillan Press: 42-51.

Perry, J. (1972). Can the Self Divide? *Journal of Philosophy*, 59, 16: 463-489.

Putnam, H. (1975). Minds and Machines in Putnam, H. *Mind, Language and Reality.* Cambridge: Cambridge University Press. 362-385.

Quine, W.V. (1969). Epistemology Naturalized in Quine, W.V. (ed), *Ontological Relativity and Other Essays.* New York: Columbia University Press: 69-90.

Rorty, R. (1979). *Philosophy and the Mirror of Nature.* Princeton, NJ: Princeton University Press.

Rumelhart, D. & McCleland, J. *Parallel Distributed Processing: Foundations.* Cambridge, Ma: MIT Press/Bradford Books.

Russell, B. (1971). *The Analysis of Mind.* London: George Allen & Unwin.

Ryle, G. (1949). *The Concept of Mind.* New York: Barnes & Noble.

Smart, J.J.C. (1989). *Our Place in the Universe.* Oxford: Basil Blackwell.

——. (1962). Sensations and Brain Processes in *The Mind/Brain Identity Theory* ed. C.V. Borst London: The Macmillan Press: 52-66.

Teixeira, J. de F. (1990). *O que é Inteligência Artificial.* S. Paulo: Editora Brasiliense.

——. (1990). *Ensaio sobre a Moral de Descartes.* S. Paulo: Editora Brasiliense.