UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS HUMANAS, LETRAS E ARTES
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

ANDRÉ ROSOLEM SANT'ANNA

O PROBLEMA EPISTÊMICO DOS *QUALIA*

MARINGÁ
2016

ANDRÉ ROSOLEM SANT'ANNA

O PROBLEMA EPISTÊMICO DOS *QUALIA*

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Filosofia do Centro de Ciências Humanas, Letras e Artes da Universidade Estadual de Maringá, como condição parcial para a obtenção do grau de Mestre em Filosofia sob a orientação da Profa. Dra. Patrícia Coradim Sita.

MARINGÁ
2016

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

S232p

Sant'Anna, André Rosolem
O problema epistêmico dos qualia / André Rosolem Sant'Anna. – - Maringá,2016.
123 f.

Orientadora: Profa. Dra. Patrícia Coradim Sita.

Dissertação (Mestrado)- Universidade Estadual de Maringá, Centro de Ciências Humanas Letras e Artes, Programa de Pós-Graduação em Filosofia, 2016.

1. Filosofia da mente. 2. Teoria do conhecimento. 3. Epistemologia. I. Sita, Patrícia Coradim, orient. II. Universidade Estadual de Maringá. Centro de Ciências Humanas, Letras e Artes. Programa de Pós-Graduação em Filosofia. III. Título.

CDD 21. ed. 128.2

MGC-001721

Aos meus pais, Sidnei e Teresa.

Agradecimentos

Aos meus pais, Sidnei e Teresa, que sempre me apoiaram em minhas decisões e sem os quais este trabalho não teria sido possível.

À Lígia, por ter lido diferentes versões do texto, pelas sugestões e pelo companheirismo ao longo do desenvolvimento deste trabalho.

Aos professores do Departamento de Filosofia da UEM pela contribuição, a partir de perspectivas filosóficas diferentes, para a minha formação em Filosofia. Agradeço especialmente ao Wagner Félix que sempre esteve disponível em sua função como coordenador do mestrado, e à Rosângela, por todo apoio nas questões práticas referentes à realização deste trabalho.

Gostaria de agradecer, em particular, à minha orientadora, Patrícia Coradim Sita, por ter orientado esse trabalho desde seu início em meu primeiro ano de graduação, e Max Rogério Vicentini, por várias discussões sobre tópicos em filosofia da mente.

Aos professores Amélia de Jesus de Oliveira e Max Rogério Vicentini, que aceitaram participar da banca de defesa e por suas sugestões ao trabalho.

À CAPES por ter financiado os meses iniciais dessa pesquisa, o que permitiu dar continuidade ao projeto.

Resumo

Um dos problemas que tem ocupado parte da agenda dos filósofos contemporâneos diz respeito à tentativa de se apresentar fundamentos sólidos para justificar nossas crenças acerca dos estados mentais de outros seres humanos ou de outros animais. Um tópico de particular relevância para essa discussão se situa na incapacidade que temos de observar diretamente os aspectos qualitativos dos estados mentais de outras pessoas e de outros seres vivos. Embora tenhamos bons motivos práticos para acreditar que compartilhamos de estados mentais de natureza semelhante, as motivações teóricas subjacentes a essa crença nem sempre são tão claras. Em outras palavras, embora não nos preocupemos com tais questões ao interagir com outras pessoas no dia a dia, elas geram problemas de particular importância para a filosofia. O meu objetivo nesta dissertação é estabelecer os fundamentos filosóficos para uma teoria epistêmica dos *qualia*. Com isso, pretendo articular noções recorrentes na literatura recente em filosofia da mente e em filosofia da biologia que permitem justificar, a partir de uma perspectiva filosófica, as nossas crenças cotidianas segundo as quais compartilhamos de estados mentais de natureza semelhante ao de outros indivíduos de nossa espécie. Inicio essa discussão apresentando como o problema do qual pretendo tratar surge na filosofia da mente, prezando por uma análise que seja sensível tanto à literatura contemporânea quanto às raízes históricas do problema. Em seguida, argumento que uma concepção contemporânea da mente — o funcionalismo — se depara com problemas de difícil solução gerados pelos aspectos qualitativos de nossos estados mentais. Proponho, como parte sintética do trabalho, a reformulação do funcionalismo a partir da aproximação deste com aspectos teóricos da biologia evolutiva. Essa aproximação permitirá a formulação do que chamo de teoria epistêmica dos *qualia*, a qual busco fundamentar filosoficamente no decorrer dessa dissertação.

Palavras-chave: qualia, funcionalismo, funções etiológicas

Abstract

One problem that occupies the agenda of many contemporary philosophers is whether we can justify our beliefs about the nature of other human beings and animals mental states. This problem arises because we cannot observe directly the mental states of other biological systems. Despite the fact that we have good practical reasons to believe that we share mental states of the same nature, the underlying theoretical motivations supporting this belief are not always so clear. In other words, although we do not take those questions into consideration when interacting with people on a daily basis, they do raise important concerns for philosophers. My aim in this dissertation is to establish the philosophical foundations for what I call an epistemic theory of qualia. In order to accomplish this goal, I bring important discussions going on in philosophy of mind and in philosophy of biology together as a means to justify our ordinary beliefs that other people and other animals have mental states with the same qualitative aspects. I start by discussing the problem with which I will be concerned, that is, the epistemic problem. I argue that a recent approach to the problem of mind-body interaction — functionalism — fails to deal with objections raised in relation to the qualitative aspects of mental states (qualia). I propose, as an alternative, a reformulation of functionalism by articulating the view in relation to some important notions in philosophy of biology, such as the notion of etiological functions. I then argue that this approximation allows us to formulate an epistemic theory of qualia which solves the epistemic problem discussed previously.

Keywords: qualia, functionalism, etiological functions

Sumário

Introdução

Os $qualia$ têm colocado um grande desafio para qualquer proposta de se abordar a mente cientificamente. Não parece possível, dado o nosso conhecimento atual, explicar como propriedades que são intrinsecamente $qualitativas$ podem ter sua origem em propriedades meramente $quantitativas$. Parece pouco provável, por exemplo, que a doçura que experimentamos ao provar o mel possa ser explicada unicamente em termos de interações eletroquímicas em nossos cérebros. Igualmente, a vermelhidão da experiência visual de uma maçã parece não se reduzir a interações neuronais no cérebro. Isso se torna mais claro quando consideramos uma situação que, embora seja pouco comum, parece ser bastante intuitiva. Imagine que pudéssemos abrir seu cérebro quando você experimenta o mel ou quando olha para uma maçã. Nesse caso, ao olharmos para as interações neuronais em seu cérebro, não veríamos nada que seja $doce$ ou $vermelho$.

Essa situação, como está implicitamente sugerido, coloca um dilema bastante interessante para aqueles que pretendem estudar a mente cientificamente. Se os $qualia$ são propriedades qualitativas das minhas experiências conscientes, e se minhas experiências conscientes não são nada mais do que processos cerebrais, então por que não "encontramos" os $qualia$ quando examinamos um cérebro? Nesse ponto, o dilema fica evidente: ou (i) negamos que os $qualia$ existem; ou (ii) reconhecemos que eles não são propriedades físicas. A menos que estejamos dispostos a negar a realidade de aspectos tão evidentes de nossas experiências conscientes, o caminho mais razoável a se seguir parece ser aceitar alguma forma de dualismo de propriedades, no qual assumimos a realidade de propriedades que não são físicas.

Não surpreendentemente, ainda no século XVII o filósofo francês René Descartes defendera que nossos estados mentais, em contraposição aos objetos corpóreos, são a única coisa de cuja existência não podemos duvidar. Para Descartes, não podemos duvidar do fato de que pensamos, visto que duvidar é, por definição, uma forma de pensar. Desse modo, pelo menos no momento em que duvido, estou seguro de que há algo do qual não posso duvidar, isto é, o fato de que duvido. Isso nos mostra, de modo bastante breve, que negar a realidade dos nossos estados mentais não consiste em uma tarefa trivial. Na verdade, se Descartes estiver correto, parece pouco provável que uma solução teórica para o problema dos $qualia$ possa se originar

em (i).

Quando Descartes estabelece que o modo de conhecimento mais seguro que podemos ter se encontra no âmbito do conteúdo de nossos estados mentais, a mente se torna um objeto de conhecimento *infalível*. Em outras palavras, embora os conteúdos dos nossos estados mentais possam não ter nenhum correspondente com uma realidade externa, eles não podem ser objeto de dúvida enquanto conteúdo dos estados mentais. Posso duvidar que vejo um carro voador agora (posso estar em um estado de alucinação forte, por exemplo), mas não posso duvidar que *penso* ver um carro voador, ou, ainda, que *parece* que vejo um carro voador.

O argumento de Descartes que estabelece a mente como fonte de conhecimento indubitável ou infalível aponta para considerações bastante interessantes sobre o dilema mencionado acima: se essa concepção estiver correta, o que parece ser bastante intuitivo, então não podemos duvidar da existência dos $qualia$, uma vez que eles são propriedades mentais por excelência. Assim, se não estivermos dispostos a sustentar (i), resta-nos somente endossar o caminho (ii)[1].

Neste ponto, o problema dos $qualia$ parece ser mais um problema *ontológico* do que *epistemológico*, como parece sugerir a tese da infalibilidade. Na verdade, se restringirmos a investigação somente a nós enquanto indivíduos, então o problema dos $qualia$ parece ser somente uma questão de decidir se os $qualia$ são propriedades físicas ou não. O problema, no entanto, adquire uma faceta epistemológica quando nos deparamos com uma crença bastante familiar que temos, isto é, que nossos familiares e amigos, e mais ainda, todos os seres humanos, parecem possuir mentes assim como nós (individualmente) acreditamos que temos uma mente.

Para tornar isso mais claro, consideremos os dois caminhos descritos acima. Se estivermos comprometidos com a ontologia das ciências naturais e tivermos que escolher entre (i) e (ii), e mais ainda, se a tese da infalibilidade estiver correta, então o único caminho parece ser abandonar esse comprometimento ontológico ou alargar o domínio de entidades que figuram em nosso quadro categorial. Isso é bastante claro quando consideramos somente o nosso caso individual enquanto detentores de uma mente. A dificuldade surge na medida em que observamos que outros seres humanos e alguns animais parecem reagir de modo bastante semelhante

[1]Descartes era, ao contrário do que sugere (ii), um dualista substancial. A minha preocupação nessa introdução não é classificar sistematicamente a filosofia de Descartes, mas somente destacar a concepção de mente como fonte de conhecimento infalível. Com isso, acredito que ambos dualismo substancial e dualismo de propriedades são compatíveis com a tese da infalibilidade, mas não precisam necessariamente se comprometer com ela.

ao que reagimos ao interagir com os objetos do mundo. Quando encosto minha mão em uma superfície quente, por exemplo, rapidamente a distancio dessa superfície. Dependendo da gravidade da queimadura, posso estremecer, emitir alguns gritos de dor ou até mesmo desmaiar. Esses comportamentos são característicos em situações nas quais não somente eu, mas também outros seres humanos e animais, sentem dor.

Embora seja razoável atribuir estados mentais a outros seres que reagem de modo significativamente semelhante ao modo em que agimos nessas situações, não podemos estar absolutamente seguros de que os mesmos estados mentais aos quais estamos sujeitos também sejam o caso nesses outros indivíduos. Isso ocorre porque eu não posso sentir a *sua* dor, mas somente observar que você tem reações similares às minhas quando me encontro em um estado de dor. Nesse sentido, os fundamentos das nossas crenças de que outros seres possuem uma mente são estritamente comportamentais, o que aponta para a aparente impossibilidade de termos acesso *direto* às experiências de outras pessoas e animais. Enquanto temos conhecimento absoluto e infalível de nossas próprias mentes, não podemos conhecer diretamente as mentes de outras pessoas. O único modo em que podemos ter algum conhecimento delas é indireto, o que exige uma pressuposição bastante forte, de um ponto de vista filosófico, para sustentar nossas crenças. Em outras palavras, não tendo certeza de que compartilhamos estados mentais da mesma natureza, assumo como princípio norteador que a presença de estados comportamentais semelhantes em situações relevantes é *suficiente* para me permitir conhecer a natureza dos seus estados mentais e dos estados mentais de outros animais em analogia ao meu caso particular.

Utilizar de um critério comportamental parece bastante intuitivo e reflete muito de nossa prática no dia a dia, mas, como veremos no Capítulo 1, esse critério enfrenta sérios desafios teóricos. Tendo consciência dessas dificuldades, alguns filósofos propuseram abandonar o critério comportamental em favor de uma aproximação àquilo que as ciências empíricas têm a nos dizer sobre a mente. Esses filósofos defendem que não precisamos nos restringir a uma análise comportamental de outros seres para atribuir mente a eles. Podemos olhar, por exemplo, para sua constituição biológica e buscar por similaridades no funcionamento dessas estruturas em nosso caso e no caso de outros indivíduos.

Isso talvez seja mais evidente no caso dos seres humanos. Se aceitarmos uma concepção

de mente compatível com a ontologia das ciência naturais, então parece bastante razoável dizer que a mente nada mais é do que processos cerebrais. Essa tese, no entanto, enfrenta dois grandes problemas. Primeiro, como vimos anteriormente, não parece possível identificar completamente mente e cérebro, uma vez que alguns aspectos essenciais da mente (os $qualia$, por exemplo) parecem ser deixados de lado. Segundo, o raciocínio que nos leva a associar estados mentais com estados cerebrais partilha dos mesmos pressupostos do raciocínio que associa estados mentais a estados comportamentais. A diferença consiste na alteração do domínio de identificação, visto que ambos estados comportamentais e estados cerebrais nos dão apenas evidências indiretas sobre a natureza dos estados mentais. De modo bastante geral, esse é o *problema epistêmico dos qualia* que tratarei nesta dissertação. Em outras palavras, como podemos justificar nossas crenças de que outros indivíduos possuem estados mentais com *qualia*?

Um comentário em relação ao problema epistêmico é necessário neste ponto. Esse problema, tal como o formulo neste trabalho, possui muitos pontos de interseção com um problema mais geral da filosofia da mente, isto é, o *problema das outras mentes*. No que se refere a este último, questiona-se sobre a possibilidade de sabermos se existem outras mentes no mundo, e mais ainda, qual a natureza dessas outras mentes. O problema epistêmico pode ser considerado como um problema subjacente ao problema das outras mentes. Em outras palavras, quando tratamos do primeiro, temos em vista as mesmas preocupações epistêmicas, mas focamos nossa análise em um aspecto específico dos estados mentais, isto é, seus aspectos qualitativos. Desse modo, trato de ambos os problemas ao longo dessa dissertação como sendo, em um sentido amplo, interdependentes. Quando nos questionarmos, por exemplo, como podemos saber se outros sistemas possuem mente no contexto da teoria epistêmica, esse quetionamento deve ser feito no contexto da discussão sobre os *qualia*, ou seja, como sabemos se outros sistemas possuem mentes com os mesmos aspectos qualitativos.

Para resumir essa breve discussão introdutória, vimos que os $qualia$ são propriedades que não parecem ser redutíveis às propriedades físicas do mundo. Isso nos coloca frente a um dilema: ou (i) negamos a existência dos $qualia$, o que parece pouco intuitivo; ou (ii) assumimos que os *qualia* não são propriedades físicas. O problema com a segunda alternativa é que se a redução dos $qualia$ a propriedades físicas for impossível, então um estudo científico da mente

parece ser bem limitado. Mais ainda, se os $qualia$ não forem propriedades físicas, qualquer tentativa de justificar nossas crenças sobre a natureza da mente de outros seres em termos físicos parece ser um projeto impossível. Nesse sentido, temos como objeto um problema genuinamente filosófico, isto é, questionamo-nos sobre a possibilidade de a ciência explicar aspectos que consideramos relevantes sobre a mente.

O ponto motivador do meu trabalho é que, pelo menos no que diz respeito ao problema epistêmico, a ciência é capaz de providenciar critérios seguros de justificação para nossas crenças sobre a natureza dos $qualia$ de outros indivíduos. Como ficará claro no Capítulo 1, não pretendo tomar partido na discussão ontológica sobre os $qualia$. Acredito que o problema epistêmico, embora intimamente relacionado com o problema ontológico, pode ser tratado de modo independente. Assim, para concluir esta introdução, apresentarei em linhas gerais a natureza da minha proposta. Essa dissertação consiste na tentativa de superar o problema epistêmico dos $qualia$. Para isso, desenvolvo o que chamo de *teoria epistêmica dos qualia*. Essa teoria terá duas formulações diferentes capazes de lidar com casos específicos discutidos na filosofia da mente. Como a minha proposta se insere numa perspectiva geral de otimismo sobre a possibilidade do estudo científico da mente, considero aqui uma concepção específica da natureza da mente (o *funcionalismo*) aliado a um dos pilares da ciência contemporânea: a *teoria da evolução*.

O Capítulo 1 serve de introdução ao tema da dissertação. Nele discorro sobre as principais tentativas de abordar a mente cientificamente no século XX. O Capítulo 2 é uma exploração sistemática do funcionalismo. Argumento que o funcionalismo supera dificuldades colocadas a teorias anteriores, mas que a noção de função adotada por versões clássicas do funcionalismo não consegue superar dificuldades epistemológicas relacionadas aos $qualia$. Concluo o capítulo propondo a reformulação do funcionalismo adotando uma nova concepção de função, isto é, a *concepção etiológica de função*, tentando mostrar como ela apresenta recursos para resolver o problema epistêmico. O Capítulo 3 consiste na investigação sistemática da noção etiológica de função e suas implicações filosóficas para o funcionalismo. Nesse capítulo, desenvolvo a noção de *explicação teleofuncional* que será essencial para lidar com o problema epistêmico e mostro como ela se relaciona com o funcionalismo e o problema dos $qualia$. Por fim, no Capítulo 4,

apresento a teoria epistêmica e mostro como ela pretende resolver os problemas levantados ao longo da dissertação.

Capítulo 1

Problema mente-corpo

1.1 Introdução ao problema mente-corpo

A filosofia da mente contemporânea tem como um de seus principais temas de investigação a relação entre a mente e o corpo. Dentro dessa temática geral, encontramos questionamentos mais específicos, tais como (i) quais são os processos que dão origem aos estados mentais; (ii) se esses estados mentais são redutíveis aos processos que lhes dão origem; e (iii) como a dimensão qualitativa das nossas experiências conscientes pode ser explicada em termos puramente científicos. Embora essa seja uma caracterização bastante geral, podemos estabelecer como ponto de partida de uma reflexão filosófica considerações que remetem a uma concepção bastante comum da noção de "mente". Essa concepção de "senso comum" (*folk*) da mente, como podemos nos referir a ela, estabelece alguns padrões para que possamos descrever de forma racional o comportamento e os processos considerados internos de outros indivíduos, sejam eles seres humanos ou animais situados em escalas menos complexas da escala evolutiva.

Um bom exemplo dessa concepção transparece no constante uso que fazemos de noções como "crenças" ou "desejos". Ter a crença de que uma garrafa de refrigerante está na geladeira e o desejo de tomar um pouco de refrigerante são maneiras de descrever a minha ação de me levantar da cadeira na qual estava sentado e me dirigir à cozinha para tomar um copo de refrigerante. Crenças e desejos podem ser classificados, segundo a concepção de senso comum da mente, como estados internos que habitam o que William James (1890) denominou de "fluxo de consciência", isto é, a sucessão de estados mentais ao quais estamos submetidos

quando estamos acordados. De um modo mais preciso, a minha crença de que há uma garrafa de refrigerante na geladeira é um estado mental que ocorre no fluxo das minhas experiências conscientes e que pode estabelecer novas direções para esse fluxo de consciência de acordo com o seu conteúdo.

Dentro da perspectiva do senso comum, portanto, a crença de que a garrafa está na geladeira e o meu desejo de beber um pouco de refrigerante são estados realizados em minha mente que, de alguma forma, possuem relação com o comportamento que apresento após esses estados terem passado pelo meu fluxo de consciência. Nesse sentido, se desejo algo e tenho a intenção de satisfazer esse desejo, naturalmente me engajo em uma série de movimentos corporais coordenados que terão como objetivo a satisfação desse desejo. Essa capacidade que tenho de utilizar meu corpo para satisfazer um desejo (algo que é pretensamente "mental") explicita uma importante característica dos estados mentais: isto é, eles possuem *eficácia causal* em relação ao corpo.

Mas afirmar que a mente é capaz de causar movimentos corporais que têm como objetivo satisfazer crenças e desejos abre uma gama de questionamentos de difícil solução. O primeiro e mais geral deles diz respeito à natureza dos estados mentais. Seriam estados mentais apenas processos físicos que ocorrem no nosso corpo? Ou há algum outro aspecto desses estados que não pode ser expressamente explicado apenas pelo estudo do cérebro? Dependendo de como respondemos a essas questões, novos problemas surgem. Dentre eles, um problema de grande importância é o de se explicar — de acordo com o modo em que concebemos a natureza dos estados mentais — como é possível a interação desses últimos com estados físicos ordinários como a movimentação do corpo. Em outras palavras, dado que o corpo seja parte do mundo físico, qual seria a natureza da mente de modo que os estados mentais possam ser capazes de interagir, de modo causal, com eventos físicos?

Lucrécio, ainda no século I a.C, já havia notado algumas dessas peculiaridades sobre a relação entre mente e corpo. Conforme aponta em *Da Natureza*, a relação entre mente e corpo torna-se clara uma vez que consideramos casos de patologia no corpo[1]. Se, por exemplo, somos vítimas de um mal estar causado pelo mau funcionamento de determinada parte do nosso corpo,

[1]Lucrécio apresenta uma série de argumentos que, segundo ele, indicam que há uma relação entre mente e corpo. Para uma exposição sistemática desses argumentos, ver *Da Natureza*, Livro III.

o fluxo de estados conscientes ao qual estávamos submetidos pode cessar por algum momento. Isso, segundo Lucrécio, seria evidência de que há uma interação entre mente e corpo, ainda que não saibamos a natureza exata dessa interação.

Embora Lucrécio tenha dedicado parte significativa de suas investigações com o objetivo de compreender essa relação, o problema da interação entre mente e corpo encontra-se de modo mais expressivo no projeto filosófico de René Descartes. Em linhas gerais, Descartes defende um dualismo ontológico que concebe o mundo a partir da existência de duas substâncias distintas: isto é, uma *substância pensante*, exclusivamente associada à alma; e uma *substância extensa*, característica dos corpos materiais. O homem, segundo Descartes, seria o único ser composto pela união das duas substâncias, destacando dentro de seu projeto a concepção comum e bastante intuitiva de que existe uma relação entre o que comumente chamamos de "mente" e de "corpo".

Embora Descartes reconheça essa dualidade, o problema da interação entre mente e corpo não é uma mera consequência dessa divisão. O problema surge com uma caracterização filosófica mais precisa na medida em que Descartes concebe as duas substâncias como ontologicamente independentes. Em outras palavras, tanto substância pensante quanto substância extensa são capazes de existirem por si só. Nessa perspectiva, o homem seria a única entidade no qual a união se dá de modo efetivo. A dificuldade, no entanto, é que nos parece muito claro que existe uma interação entre as substâncias, mas se ambas são inteiramente independentes uma da outra, então essa relação parece se fundar num pressuposto pretensamente misterioso.

Essa dificuldade é explorada de modo mais sistemático na ampla crítica que os filósofos da mente do século XX fazem ao projeto cartesiano. Não querendo discutir os méritos dessas críticas nesse momento — e tampouco se se fundamentam em uma leitura fiel da filosofia de Descartes — uma objeção comum tem sido a de que conceber a mente e o corpo como duas substâncias ontologicamente distintas, que por algum motivo se encontram unidas nos homens, parece contrastar com um valor bastante prezado pelas ciências naturais. De acordo com essa crítica, ao assumir que mente e corpo — ontologicamente distintos — podem interagir causalmente, estaríamos violando o princípio segundo o qual todo fenômeno do universo físico deve ter uma causa suficientemente física e, mais ainda, que essa relação causal deve ser explicada

pelas leis da física[2]. Podemos dizer que essa concepção do mundo físico aceita como ponto de partida a tese segundo a qual o universo físico está *causalmente fechado*.

> **Tese do Fechamento Causal do Mundo Físico**: Todo fenômeno físico *F* deve ser explicado *necessariamente* e *suficientemente* por uma causa de natureza física C_f.

Para tornar esse ponto mais claro, consideremos um exemplo. Se, como pressupõe a tese do fechamento causal do universo físico, todo evento físico possui uma causa necessariamente e suficientemente física, segue-se disso que a minha ação de me dirigir à cozinha para beber um copo de refrigerante deve ter causas suficientemente físicas. Isso, no entanto, parece inviabilizar qualquer explicação que atribua poder causal às minhas crenças e desejos, visto que, em uma concepção que preza pela separação ontológica entre mente e corpo, o meu desejo de tomar um copo de refrigerante enquanto um estado mental é um estado de natureza não-física que exerce influência causal sobre o meu corpo. Nesse sentido, o meu comportamento de me dirigir à cozinha e pegar a garrafa teria causas físicas, a saber, a operação dos neurônios em meu cérebro, e causas não-físicas, como o desejo de beber um pouco de refrigerante. Essa "dupla causação" do meu movimento de me dirigir à cozinha para beber refrigerante viola o princípio de fechamento causal do mundo físico, uma vez que a postulação de estados mentais de natureza não-física como causa de eventos físicos parece estabelecer uma *sobredeterminação causal* sobre o evento físico. Nessa perspectiva, eventos físicos teriam causas mentais além de suas causas físicas usuais. Esse argumento pode ser esquematizado da seguinte forma:

> (P1) Movimentos corporais m são eventos físicos;
> (P2) m é causado por causas físicas anteriores;
> (P3) Ou m tem causas suficientemente físicas ou m tem causas suficientemente mentais;
>
> Portanto,
>
> (C) Não é o caso que m tem causas de natureza mental.

Frente a essa dificuldade, e não satisfeitos com uma abordagem que preza explicitamente por uma divisão ontológica entre mente e corpo, alguns filósofos da mente contemporâneos

[2]Defendendo uma forma de naturalismo que toma como ponto de partida essa ideia, Papineau escreve: "Considero que a física, em contrapartida a outras ciências especiais, é completa no sentido em que todos os eventos físicos são determinados, ou têm suas chances determinadas, por eventos físicos prioritários regidos por leis físicas". (PAPINEAU, 1993, p. 16)

propuseram, assim como Lucrécio, que a mente seria essencialmente física, e, portanto, apenas mais um elemento do mundo físico. Isso resguardaria, em princípio, a capacidade de a mente interagir causalmente com o corpo, evitando a sobredeterminação causal, e, por consequência, a violação do princípio do fechamento causal do universo físico. Assim, baseados em evidências aduzidas por teorias recorrentes na neurociência, alguns filósofos propuseram a tese de que o cérebro seria a estrutura física responsável pela realização dos estados mentais. De uma forma geral, atribuir ao cérebro a localização espacial onde se situa a mente significa dizer que crenças, desejos, etc., poderiam ser explicados a partir do estudo das interações eletroquímicas ocorrentes entre os bilhões de neurônios que compõem nosso cérebro.

Essa concepção *monista* da relação entre mente e cérebro, como podemos chamá-la em oposição ao dualismo cartesiano, enfrenta algumas dificuldades próprias em relação a alguns aspectos considerados essenciais no estudo da mente. Considere, por exemplo, o caso em que tomo um copo de refrigerante. Ao tomar o refrigerante, tenho determinada *sensação* como, por exemplo, a *doçura* do líquido e uma leve *ardência* por causa de sua gaseificação. Assim, quando tomo o refrigerante, tenho a experiência consciente da doçura, isto é, tenho a experiência consciente de *como é experimentar* algo doce. Estados mentais dessa natureza possuem aspectos qualitativos como os de "ser doce" ou "ser ardente". Nessa perspectiva, uma teoria que pretenda localizar o cérebro como o cerne da mente deve ser capaz de explicar como os estados mentais com aspectos qualitativos são possíveis. Conforme mencionado acima, entretanto, os estados mentais com aspectos qualitativos parecem oferecer dificuldades para que uma explicação da mente que segue os pressupostos de uma postura fisicalista seja possível. Isso ocorre, em parte, porque o mundo físico, do qual os processos biológicos fazem parte, parece ser alheio a qualquer caracterização *qualitativa*, restringindo sua constituição a aspectos unicamente *quantitativos*. Na literatura contemporânea, é comum se referir a esse problema como o "problema dos *qualia*", no qual *qualia* (singular *quale*) se referem aos aspectos qualitativos de nossos estados mentais.

Para entendermos o problema dos *qualia* de um modo mais detalhado, considere a dor que sinto ao espetar o dedo em uma agulha. A dor enquanto aspecto qualitativo de um estado mental é considerada como um estado intermediário entre determinados *inputs* ambien-

tais, como dano a algum tecido do corpo, e determinados *outputs* comportamentais, como um gemido e a tentativa de aliviar a dor. Assumindo que a mente esteja situada no cérebro, a experiência da dor enquanto estado mental deve ser realizada no cérebro, o que implica que seus aspectos fenomenais e qualitativos devem ser, de algum modo, explicados pelas interações eletroquímicas realizadas entre neurônios. É neste ponto que o problema dos $qualia$ se torna expressivo: parece não haver qualquer indício de como os aspectos qualitativos de nossas experiências conscientes têm sua origem a partir do funcionamento dos neurônios nas mais variadas interações realizadas no cérebro. Em outras palavras, não há nenhuma relação necessária entre determinada interação eletroquímica realizada por neurônios em determinada parte do cérebro e o aspecto qualitativo da experiência consciente relacionada a essas interações. Nesse sentido, a relação entre processos físicos no cérebro e aspectos qualitativos dos estados mentais parece ser uma relação totalmente contingente, uma vez que parece não haver um fato físico que determine o porquê de determinados *inputs* ambientais e *outputs* ambientais serem acompanhados por um aspecto qualitativo *q* em vez de um aspecto qualitativo *r*, ou seja, não há nenhum fato físico que evidencie por que os *inputs* e *outputs* relacionados à dor são acompanhados pela sensação de dor e não pela sensação de cócegas, por exemplo.

Os estados mentais com *qualia*, no entanto, têm outro aspecto problemático. O fato de a experiência de dor não poder ser evidenciada pela observação da interação de neurônios em nosso cérebro parece ameaçar a possibilidade de comunicação interpessoal em relação a esses fenômenos, uma vez que não podemos saber de fato se as outras pessoas realmente experienciam a mesma sensação de dor que nós experienciamos. Neste ponto, a observação do comportamento enquanto evidência do caráter intersubjetivo do aspecto qualitativo da dor não apresenta nenhuma vantagem: não há nenhum fato relativo aos movimentos do corpo que indique uma relação necessária entre o comportamento e o aspecto qualitativo da dor. Nesse sentido, o aspecto qualitativo da dor parece ser inefável, isto é, somente acessível ao sujeito da qual ela é dor, o que cria o ensejo para abordagens céticas em relação à natureza dos *qualia*.

Tendo em vista as considerações acima, as pressuposições básicas de uma teoria fisicalista da mente parecem enfrentar dificuldades, assim como o dualismo cartesiano, para acomodar os fenômenos mentais em sua completude dentro de seu aparato teórico. Analisarei, nas

seções seguintes, como se desenvolveram as teorias de orientação fisicalista sobre o problema mente-corpo no século XX, apontando para as suas subsequentes dificuldades. Essa discussão servirá como pano de fundo conceitual para a abordagem do problema epistêmico dos *qualia*.

1.2 O problema mente-corpo no século XX

1.2.1 Behaviorismo

Caracterizado de modo amplo, o *behaviorismo lógico* ou *behaviorismo filosófico* consiste na tentativa de estudar os fenômenos mentais a partir do comportamento. Nesse sentido, estados mentais como a dor seriam estados de natureza interna — isto é, somente acessíveis ao indivíduo — capazes de causar determinados movimentos físicos no sistema do qual ele faz parte. Esses movimentos dos quais estados mentais são causa são subsequentemente caracterizados como o "comportamento" exibido por esse sistema na presença de um estado mental específico.

Como ponto de partida, é importante ter em mente que o behaviorismo lógico não implica, necessariamente, o behaviorismo psicológico clássico associado a autores como Watson e Skinner. O behaviorismo lógico (BL, a partir de agora) é uma tese epistêmica que diz respeito à caracterização e ao estabelecimento das relações entre estados mentais e estados comportamentais. Nesse sentido, podemos dizer que o BL consiste em uma tese mais fraca que as tradicionais versões do behaviorismo psicológico.

Mas no que exatamente consiste o BL? Como vimos acima, o BL é primariamente uma *tese lógico-epistêmica*. Por "tese lógico-epistêmica" podemos entender uma tese que busca elencar e estabelecer relações entre estados mentais e estados comportamentais tal que, dado um conjunto C de estados comportamentais do tipo $c_1 \dots c_n$, podemos *saber* com segurança que um conjunto M de estados mentais do tipo $m_1 \dots m_n$ são o caso em um sistema *S*. Em termos lógicos, temos que:

> **(i)** se um estado mental m_1 é o caso, então seu correlato comportamental c_1 (ou as disposições para realizer c_1) é necessariamente o caso;
>
> **(ii)** se um comportamento c_1 (ou a disposição para realizer c_1) é o caso, então seu correlato mental m_1 é necessariamente o caso;

O que nos leva à tese mais geral:

(iii) $M \leftrightarrow B$ (Estado mental *se e somente se* estado comportamental ou disposição para realizar estado comportamental)

Nessa perspectiva, é seguro dizer que embora o BL pretenda estudar os fenômenos mentais a partir da observação do comportamento de um determinado sistema, seus pressupostos filosóficos não implicam diretamente uma tese ontológica sobre a relação entre o estado interno de se ter uma dor e o comportamento causado por essa dor. Em outras palavras, o BL não pressupõe que estados mentais sejam *identificados* com seus *outputs* comportamentais, pretendendo somente descrever o modo pelo qual apreendemos e descrevemos esses estados mentais. O BL pode ser considerado, portanto, como uma perspectiva teórica que descreve o modo pelo qual observamos e estabelecemos relações entre os estados mentais. Daí sua caracterização como uma tese lógico-epistêmica.

Uma vantagem importante que o BL adquire ao assumir que estados comportamentais revelam a presença de estados mentais, e mais ainda, que uma correlação pode ser estabelecida entre esses dois domínios, reside no fato de que sua proposta apresenta perspectivas para um estudo da mente que pode superar sua aparente *inefabilidade*. A noção de "inefabilidade", bastante importante para os nossos propósitos, diz respeito à impossibilidade de conhecermos *diretamente* os estados mentais de outras pessoas. Embora tenhamos evidências indiretas que nossos familiares e amigos possuam estados mentais semelhantes aos nossos, nunca temos acesso direto aos seus estados mentais propriamente ditos.

Essa assimetria é expressa de um modo bastante preciso na distinção que podemos fazer entre duas perspectivas a partir das quais podemos julgar sobre a natureza das coisas: a *perspectiva de primeira pessoa* e a *perspectiva de terceira pessoa*. De um modo geral, o conhecimento gerado a partir da perspectiva de primeira pessoa é dado a partir da perspectiva única de um sujeito cognoscente. Em termos mais familiares, isso pode ser expresso pela aparente capacidade que temos de tomar como objetos de conhecimento os nossos próprios estados mentais a partir de uma suposta faculdade de introspecção. O que torna esse tipo de conhecimento característico é o fato de termos um acesso *privilegiado* a eles, isto é, um estado mental de natureza m_s só pode ser objeto de conhecimento direto para o sistema *S* ao qual ele pertence. Outros

sistemas S_1 e S_2, que não partilham de m_s, só podem ter conhecimento indireto de m_s. Em contraposição ao conhecimento de primeira pessoa, podemos destacar o tipo de conhecimento que é acessível a partir de uma perspectiva de *terceira pessoa*. Como é de se esperar, um conhecimento acessível a partir da perspectiva de terceira pessoa é aquele conhecimento que pode ser considerado e validado por diferentes sujeitos cognoscentes. Podemos dizer que o conhecimento em terceira pessoa é de natureza *intersubjetiva*, enquanto que o de primeira pessoa é essencialmente *subjetivo*.

A grande vantagem que o BL traz para o cenário da discussão da relação mente-corpo é que a mente parece ser, finalmente, passível de um estudo em terceira pessoa. Isso ocorre porque o BL assume como ponto de partida uma correlação lógica entre o domínio dos estados mentais com o domínio dos estados comportamentais. Assim, se estados comportamentais são intersubjetivamente acessíveis a diferentes sujeitos cognoscentes, e se há uma correlação lógica entre estados mentais e estados comportamentais, então estudar o comportamento implica estudar a mente. O que torna isso uma grande vantagem para o BL é o fato de que a mente passa a ser objeto direto de estudo científico, uma vez que a ciência demanda conhecimentos que sejam intersubjetivamente acessíveis. Além disso, a associação lógica entre estados mentais e estados comportamentais assegura que sentenças linguísticas que se referem a estados mentais possam ter seu conteúdo compartilhado.

Outro conceito importante para o BL, além de sua ampla fundamentação na noção de comportamento, é o de *propensão*. Como destaca Jaegwon Kim (1998), de acordo com o BL, "ter uma mente é apenas questão de se exibir, ou ter *propensão* a se exibir, certos padrões apropriados de comportamentos observáveis" (KIM, 1998, p. 26, itálicos meus). Mas o que exatamente significa ter propensões para se exibir determinados padrões comportamentais observáveis?

O BL, conforme mostra Kim (ver KIM, 1998, pp. 29-31), assume que os aspectos semânticos dos estados mentais podem ser traduzidos em termos comportamentais sem nenhuma perda, o que, naturalmente, exige por parte do BL uma explicação de como esse tipo de tradução pode ser feita. Nesse contexto, a noção de "propensão" para se exibir determinados padrões comportamentais consiste justamente na tentativa de se responder a esse problema. Em

outras palavras, estar propenso a se comportar de determinado modo *C* significa que, dado a realização de condições adequadas *A*, o comportamento *C* em questão será realizado por um sujeito *S*. Como ilustração, considere a crença de João na existência de extraterrestres. Mesmo que em um instante *t* João não apresente nenhum comportamento que permita ao behaviorista traduzir o aspecto semântico de sua crença, ele ainda assim pode afirmar que João acredita na existência de extraterrestres a partir da consideração da *propensão* ou *disposição* de João para exibir comportamentos coerentes com a sua crença de que extraterrestres existem. Desse modo, se questionado sobre a existência de extraterrestres (o que consiste numa condição adequada para que João se comporte de modo que sua crença seja exibida em termos comportamentais), João revelará sua crença na existência desses últimos.

Ao considerar o caráter disposicional do comportamento, o BL permite uma análise contrafactual dos estados mentais. Assim, um teórico que aceita o BL pode analisar estados mentais como crenças sem que esses estados sejam o caso atualmente. Embora esse artifício teórico pareça bastante propício, ele trará dificuldades insuperáveis ao behaviorismo que exploraremos com mais detalhe no próximo tópico.

Objeções ao behaviorismo

A tentativa do BL de analisar os estados mentais a partir da disposição de comportamentos em circunstâncias adequadas enfrenta algumas dificuldades que ameaçam a proposta geral do behaviorismo. Explorarei essas dificuldades com mais detalhes abaixo.

Análise disposicional inifinita

A primeira dificuldade enfrentada pelo BL diz respeito à possibilidade de atribuição infinita de condicionais em uma análise disposicional do comportamento. Conforme destacado no exemplo acima, o comportamento verbal de João só é realizado na medida em que determinadas circunstâncias são satisfeitas. O problema dessa análise reside no fato de que o conjunto de condicionais classificados como "circunstâncias adequadas" para se analisar a crença de João

pode ser aumentado infinitamente. Considerando ainda o exemplo de João, se questionado sobre a existência de extraterrestres, João expressará sua crença somente se (i) desejar dizer a verdade, visto que (ii) ele não quer que riam dele, mas (iii) ele só desejará dizer a verdade caso seu interlocutor seja uma mulher, uma vez que (iv) João acredita que as mulheres são mais compreensíveis em relação a esse assunto, e assim sucessivamente. A atribuição infinita de condicionais tem como consequência o fato de que estados mentais como crenças e desejos não podem ser definidos sem fazer referência a outros estados mentais, o que gera uma circularidade na análise disposicional do comportamento. Outra dificuldade importante a se notar em relação à atribuição disposicional de estados mentais diz respeito à impossibilidade prática de se descrever o conteúdo dos estados mentais de um determinado sujeito, uma vez que estados mentais, no perspectiva do BL, não podem ser definidos a não ser pela referência a outros estados mentais.

Contingência ontológica

Uma segunda dificuldade imposta ao behaviorismo reside no fraco vínculo ontológico estabelecido entre estados mentais e estados comportamentais. Hilary Putnam (1968) sugere que imaginemos uma sociedade na qual os indivíduos são treinados de tal maneira que, ao sentirem uma dor, continuam a agir normalmente como se essa dor simplesmente não tivesse ocorrido. Esses indivíduos, os quais Putnam denomina de *super-espartanos*, não estremecem e nem emitem qualquer ruído quando sentem uma dor. Para Putnam, esse cenário mostra que, caso o BL esteja correto, então não se pode atribuir o estado mental de dor aos super-espartanos. A dificuldade que surge diz respeito ao fato de que embora o super-espartano não exiba nenhum sinal externo de dor, ainda assim parece ser muito evidente que ele tem a sensação de dor. Na perspectiva do BL, entretanto, o indivíduo não estaria sentindo dor, visto que, uma resposta positiva à questão recorreria à análise disposicional. Mas, conforme vimos acima, esta última enfrenta dificuldades como o problema da circularidade ou da atribuição infinita de condicionais.

Contingência epistêmica

Outra objeção feita ao BL diz respeito ao caráter contingente da relação entre os estados mentais "internos" e os estados comportamentais externos. Dado que um determinado sistema exibe estados físicos comportamentais relacionados a um estado interno de dor, não parece haver nenhum elemento nessa correlação que indique uma conexão necessária entre comportamento e dor. Mais especificamente, o fato de o sistema estremecer e emitir alguns gemidos ao espetar seu dedo em uma agulha não nos permite concluir que esse sistema tenha um estado interno tal como o estado de sentir uma dor. Nesse sentido, um indivíduo A que realiza um estado interno qualitativo q e exibe comportamentos do tipo c seria, do ponto de vista descritivo assumido pelo BL, indistinguível de um indivíduo B que realiza um estado qualitativo r e exibe comportamentos do tipo c. Isso resultaria em um cenário no qual A e B possuem os mesmos estados comportamentais, mas diferem em seus estados internos.

Chauvinismo especista

Uma última objeção a ser considerada nessa exposição está relacionada a uma espécie de "chauvinismo especista" pressuposta pela relação entre um estado mental e um estado comportamental estabelecida pelo BL. Colocando em foco o exemplo da dor, a descrição comportamental desta última parece ser sensível somente ao comportamento particular de uma determinada espécie ou um determinado grupo de espécies. Podemos imaginar, por exemplo, marcianos que, a despeito de terem o mesmo estado qualitativo de dor ao tocarem em uma panela quente, ainda assim exibem estados comportamentais totalmente distintos dos que observamos em seres humanos e em algumas espécies de animais. Nesse caso, teríamos que dizer que os marcianos não estão em um estado de dor, mas em outro estado qualquer.

Tendo em vista as considerações apresentadas até aqui, o BL parece não apresentar critérios bem definidos que possam captar a complexidade associada aos fenômenos mentais. Como mostra Putnam, a proposta behaviorista parece aludir a consequências pouco intuitivas — como no caso do super-espartano — e, conforme vemos no exemplo de João, a definição de

estados mentais em termos de estados comportamentais parece não ser uma definição adequada, visto que a própria definição de um estado mental faz referência a um outro estado mental, o que gera uma definição circular. Mais ainda, o aspecto restrito da análise de estados mentais em termos de comportamento parece restringir a realização dos primeiros somente a algumas espécies de animais, uma vez que a relação entre um estado mental qualitativo e estados comportamentais não é uma relação necessária, mas apenas contingente.

1.2.2 Teoria da identidade

Uma alternative ao BL é a teoria da identidade, que foi fortemente motivada pelas crescentes relações observadas entre o funcionamento da mente e o funcionamento do cérebro. Assim como o behaviorismo, a teoria da identidade pretende evitar a intrincada distinção entre os diferentes modos de acesso epistemológico que temos ao domínio do físico e ao domínio do mental, de modo que sua proposta consiste na tentativa de situar o estudo dos estados mentais em uma perspectiva objetiva ou de terceira pessoa.

Conforme vimos anteriormente, ainda na antiguidade Lucrécio notara que estados mentais possuem uma relação evidente com o funcionamento do nosso corpo. Guiados pelos avanços da neurociência na primeira metade do século XX, teóricos como J. J. C. Smart (1959) e U. T. Place (1956) propuseram a identificação de estados mentais com estados neurofisiológicos realizados no cérebro humano[3]. Uma das grandes motivações da teoria da identidade pode ser encontrada na tentativa de evitar aquilo que chamamos de *sobredeterminação causal*. Em outras palavras, uma teoria que consiga identificar os estados mentais com estados cerebrais não violará o princípio de fechamento causal do mundo físico uma vez que, se estados mentais são idênticos a estados cerebrais, então o meu desejo de levantar o braço, por exemplo, enquanto um estado cerebral, tem causas suficientemente físicas.

Tendo em vista as objeções feitas ao behaviorismo em relação à contingência da relação entre um estado mental e um estado comportamental, os teóricos da identidade afirmam que a

[3] "Essa posição, formulada e explicitamente defendida como uma solução para o problema mente-corpo no final dos anos 1950, sustenta que estados mentais podem ser identificados com processos físicos no cérebro. Assim como não existem relâmpagos como fenômenos distintos de cargas elétricas atmosféricas, não existem estados mentais distintos dos eventos neurais (que são, em última instância, psico-químicos) que acontecem no cérebro." (KIM, 1998, p. 52)

redução de estados mentais a estados cerebrais é uma questão empírica, porém necessária. Isso quer dizer que não há nada no conceito de dor ou de estimulação de fibras-C que possibilite estabelecer a redução de modo *a priori*, mas uma vez que a redução é estabelecida de modo empírico, ela se torna necessária. A identidade é, desse modo, apenas contingente no sentido em que o teórico da identidade não se compromete em fornecer um argumento *a priori* para a identificação de um estado mental e seus estados cerebrais correspondentes.

De acordo com Churchland (1984), o tipo de redução pressuposta pelos teóricos da identidade é um redução *interteórica*, isto é, os teóricos da identidade acreditam ser possível reduzir uma teoria T_1 (supostamente sobre a natureza da mente) e seus pressupostos a uma teoria T_2 (que diz respeito ao funcionamento do cérebro). De acordo com essa perspectiva, a proposta de redução dos teóricos da identidade consiste na tradução de relações postuladas por uma teoria escolhida como ponto de partida (T_1) em termos de relações postuladas pela teoria de destino da redução (T_2).

É importante ressaltar, no entanto, que esse tipo de redução assume que a teoria de destino seja capaz de captar as peculiaridades preditivas e explicativas que estão pressupostas na teoria de partida. Para ver isso mais claramente, considere uma teoria de partida T_1 e uma teoria de destino T_2 quaisquer. Para que uma redução interteórica tal como a descrita por Churchland (1984) seja possível, os fenômenos explicados e as capacidades preditivas de T_1 devem ser explicados por T_2, e, ainda, igualmente pressupostos em suas capacidades preditivas[4]. Assim, podemos dizer que a teoria da identidade pressupõe que todos os fenômenos explicados por uma teoria da mente podem ser explicados e preditos por uma teoria que se paute nos eventos físicos realizados no cérebro. Se esse não for o caso, os propósitos de redução não seriam adequados, uma vez que o processo de redução ocasionaria a perda da capacidade explicativa e/ou preditiva de determinados fenômenos do mundo.

Objeções à teoria da identidade

[4]É importante notar que as capacidades preditivas de T_2 não precisam ser iguais às capacidades preditivas de T_1. O único requisito é que T_2 seja capaz de incorporar todas as capacidades preditivas de T_1, o que não quer dizer que as duas teorias têm as mesmas capacidades preditivas. Esse é um dos pontos importantes de uma redução, a saber, a incorporação de teorias com menor poder preditivo em teorias com maior poder preditivo.

Para melhor entender a teoria da identidade, considerarei nesta seção três tipos de objeções à proposta de redução dos estados mentais a estados cerebrais: (i) objeções relacionadas à lei de Leibniz, (ii) a objeção dos "designadores rígidos" de Kripke; e (iii) o argumento da múltipla realização.

A lei de Leibniz

A lei de Leibniz, como destaca Churchland (1984), afirma que "dois itens são numericamente idênticos somente no caso em que qualquer propriedade possuída por um dos itens também seja possuída pelo outro item." (CHURCHLAND, 1984, p. 29). Assim, segundo essa formulação da lei de Leibniz, se os estados mentais forem passíveis de redução a estados cerebrais, as propriedades dos primeiros devem ser igualmente passíveis de redução às propriedades dos últimos. Em termos formais, isso quer dizer que:

Se

(i) $M \equiv F$

Então

(ii) $\forall x(Mx \equiv Fx)$

A primeira objeção baseada nos pressupostos da lei de Leibniz diz respeito aos *aspectos semânticos* dos estados mentais. Discussões sobre a natureza de estados mentais como crenças e desejos têm seu foco no *conteúdo proposicional* associado a esses estados[5]. A minha crença de que a Terra gira em torno do Sol pode ser resumida como uma crença em uma proposição, a saber, "A Terra gira em torno do Sol". Assim, dado que o conteúdo da minha crença se resume a uma proposição, a teoria da identidade violaria a lei de Leibniz ao afirmar que estados mentais são idênticos a estados cerebrais, uma vez que estados cerebrais não parecem apresentar nenhum tipo de conteúdo proposicional. Em outras palavras, ter conteúdo proposicional seria algo exclusivo de um estado mental e que não é refletido na relações estabelecidas pelos estados cerebrais. A dificuldade que surge a partir dessas considerações é que um aspecto importante

[5] Ver Searle (1983) para uma discussão sobre o assunto.

dos estados mentais (seu conteúdo proposicional) não é passível de redução a estados cerebrais, o que viola os critérios de identidade que respeitam a lei de Leibniz. Outra objeção nessa mesma linha seria a de que estados cerebrais possuem localização espacial, enquanto estados mentais não a possuem. Se isso for o caso, no entanto, a identidade não seria possível já que um aspecto importante dos estados mentais não seria reduzido a aspectos dos estados cerebrais.

Uma segunda objeção relacionada à lei de Leibniz parte de pressupostos semelhantes aos da objeção relacionada aos aspectos semânticos e temporais dos estados mentais. Segundo essa objeção, os aspectos fenomenais da mente, como a experiência da cor vermelha, não podem ser identificados com estados cerebrais, visto que não é possível se observar nenhum estado cerebral da cor vermelha. Nesse sentido, os aspectos qualitativos dos estados mentais não seriam redutíveis aos estados cerebrais, o que violaria a lei de Leibniz no mesmo sentido da objeção discutida acima[6].

Em termos formais, isso quer dizer que se um estado mental *m* é *v* (vermelho), então, de acordo com a teoria da identidade, *m* pode ser identificado com um estado físico *f* que também é *v*. O problema é que parece ser possível que V_m ("m é vermelho") seja o caso sem que V_f ("f é vermelho") seja necessariamente o caso. Isso, no entanto, viola a lei de Leibniz, uma vez que há um elemento no domínio da teoria *M* que não tem correspondente no domínio da teoria *F*. Assim, há pelo menos um caso em que *x* é *M* e *x* não é *F*, o que implica a seguinte sentença:

(iii) $\exists$x(Mx $\wedge$ $\neg$Fx)

Em termos lógicos, portanto, temos que a teoria da identidade implica uma sentença (iii) que contradiz os pressupostos da lei de Leibniz expressos por (i) e (ii). Nesse sentido, se a noção de identidade proposta pelos teóricos da identidade for uma noção de identidade leibniziana, então ela não é capaz de reduzir efetivamente estados mentais a estados cerebrais.

[6]Em "Is Consciousness a Brain Process?", Place responde a essa objeção dizendo que ela comete o que ele denomina de "falácia fenomenológica". Para Place, o tipo de identidade pressuposta pela teoria da identidade não sustenta que os aspectos fenomenológicos (ou qualitativos) de nossos estados mentais sejam identificados com estados cerebrais, isto é, a vermelhidão da experiência visual de uma maçã não é o objeto da redução. De acordo com esse ponto de vista, o neurocientista deve somente explicar os processos cerebrais que ocorrem quando olhamos para uma maçã e não como a cor vermelha pode ser traduzida em termos desses processo cerebrais. Para mais sobre essa discussão, ver Place, (1956), pp. 58-60.

Os "designadores rígidos" de Kripke

Em *Naming and Necessity*, Saul Kripke (1980) apresenta um argumento contra a tese segundo a qual a identidade entre estados mentais e estados cerebrais é uma identidade capaz de ser estabelecida empiricamente. Segundo Kripke, para que uma identidade empírica seja possível é preciso que pelo menos um dos lados da identidade seja um *designador não-rígido*. Para ver isso de modo mais claro, considere a seguinte proposição: "O primeiro presidente do Brasil foi Deodoro da Fonseca". A relação entre o primeiro lado da identidade, isto é, "o primeiro presidente do Brasil", e o segundo lado, "Deodoro da Fonseca", é uma identidade empírica, uma vez que o fato de Deodoro da Fonseca ter sido o primeiro presidente do Brasil é apenas uma contingência histórica. Em outras palavras, é plenamente concebível que em um mundo semelhante ao nosso Deodoro da Fonseca não tenha sido o primeiro presidente do Brasil. Nesse sentido, essa identidade só é contingente porque um dos lados, de acordo com Kripke, é um designador não-rígido; mais especificamente, o primeiro lado da identidade, "o primeiro presidente do Brasil", visto que essa última expressão pode se referir a diferentes pessoas em mundos distintos.

Kripke afirma, no entanto, que no caso da identidade entre estados mentais e estados cerebrais, ambos os lados da identidade são designadores rígidos. Retomando o exemplo acima, o nome próprio "Deodoro da Fonseca" é um designador rígido no sentido em que parece não ser concebível a existência de um mundo no qual Deodoro da Fonseca não seja Deodoro da Fonseca enquanto indivíduo, ainda que este não tenha sido o primeiro presidente do Brasil. Assim, Kripke sustenta que a identificação de um estado mental como a dor, com um estado físico, como a estimulação de fibras-C, parece pressupor uma identidade na qual ambos os lados são designadores rígidos. Em outras palavras, não parece ser concebível que a dor enquanto dor possa ser de outra maneira em outro mundo possível. Caso fosse de outra maneira, não hesitaríamos em dizer que não é mais uma dor, mas algo distinto. Igualmente no caso das fibras-C, não parece plausível afirmar que um evento que é uma estimulação de fibras-C não tenha sido deste modo em um mundo possível, isto é, não parece coerente afirmar que em um outro mundo uma estimulação de fibras-C não seja uma estimulação de fibras-C. Isso implicaria a violação

do princípio da contradição, uma vez que diríamos que algo é e não é ao mesmo tempo.

Tendo estas considerações em vista, o argumento de Kripke parece indicar que a teoria da identidade apresenta uma contradição na afirmação de que a identidade entre estados mentais e estados cerebrais seja uma identidade contingente. Afirmar, todavia, que a relação seja necessária ocasionaria dificuldades semelhantes às dificuldades propostas ao behaviorismo, visto que nada nos permite afirmar que a relação entre ambos os estados seja uma relação necessária. Nesse sentido, o argumento de Kripke parece indicar que seja qual for o caminho escolhido pela teoria da identidade, ambos serão problemáticos.

A múltipla realização dos estados mentais

A teoria da identidade parece sofrer do mesmo tipo de "chauvinismo especista" atribuído ao behaviorismo, uma vez que adota uma abordagem consideravelmente restrita dos estados mentais. Em outras palavras, a teoria da identidade pressupõe que um estado mental como a dor seja identificado com determinados padrões cerebrais como a estimulação de fibras-C, o que traz como consequência o fato de que um organismo ou sistema só pode realizar um estado de dor se e somente se houver a estimulação de fibras-C nesse organismo. Essa afirmação parece excluir a possibilidade da ocorrência de dor em algumas espécies de animais, assim como em sistemas não-orgânicos. Nesse sentido, animais que se situam em um patamar inferior da escala evolutiva não poderiam realizar o estado de dor, o que parece pouco intuitivo.

A teoria da identidade, tal como formulada aqui, parece implicar que os estados mentais estão restritos a um determinado grupo de organismos capazes de realizar determinados estados físicos, o que traz como consequência o fato de que a mente é um processo restrito a organismos com determinada complexidade. Mas isso não parece adequado, visto que animais não constituintes deste grupo parecem ser capazes de realizar estados mentais como a dor, por exemplo, e os avanços em áreas como a Inteligência Artificial parecem não descartar a possibilidade da criação de sistemas não-orgânicos que possuam uma mente. Desse modo, a teoria da identidade parece ser muito restrita quando tem que lidar com estados mentais ou com a possibilidades desses estados em outros seres ou sistemas não-humanos.

Uma proposta alternativa a essa dificuldade apresentada pela teoria da identidade consiste na tese da "múltipla realização dos estados mentais" proposta por Putnam (1967), na qual estados mentais são identificados com o seu papel causal dentro de um organismo, sem fazer referência ao meio material em que esse estado é realizado. A proposta de Putnam dará origem às abordagens funcionalistas dos estados mentais que serão de importância central para essa dissertação.

1.2.3 Funcionalismo

Tendo em vista os problemas enfrentados pela teoria da identidade, Hilary Putnam (1967) propõe que pensemos em estados mentais como *estados funcionais* que possuem um determinado *papel causal* no funcionamento de um organismo. Ao propor essa tese, Putnam argumenta em favor da tese da "múltipla realização" (MR) dos estados mentais. De acordo com a MR, a identidade de uma dor com um estado físico não é mais dada em função de estados comportamentais ou estados cerebrais, mas sim em termos do papel causal exercido pelo estado mental dentro de um organismo. Ao apelar para uma definição funcional dos estados mentais, Putnam consegue evitar que sua concepção estabeleça de antemão o meio em que essa relação causal deve ser realizada.

A tese da múltipla realização dos estados mentais proposta por Putnam resulta em uma forma mais "fraca" de teoria da identidade que podemos chamar de "teoria da identidade de *ocorrências*" em contraposição a "teoria da identidade de $tipos$". Esta última diz respeito à teoria da identidade tal como tratada na seção anterior, visto que, segundo essa proposta, "tipos" de estados mentais devem ser identificados com "tipos" de estados físicos. Em outras palavras, essa perspectiva pressupõe que a identidade ocorre em um nível abstrato e não no nível das ocorrências particulares de eventos mentais e eventos físicos.

A teoria de ocorrências, por outro lado, assume que um estado mental do tipo *M* pode ser identificado com múltiplas "ocorrências" físicas $F_1 \ldots F_n$. Nesse caso, o que define o caráter de *M* é o papel causal exercido por ele dentro de um determinado sistema *S* no qual ele é realizado. O funcionalismo, desse modo, compromete-se com uma tese mais fraca de identidade, embora não admita de fato uma identidade em termos de redução de tipos. Podemos dizer, de modo

mais formal, que

> **Identidade de ocorrências**: para que um estado mental do tipo *M* seja identificado com um estado físico do tipo *F*, basta que uma das ocorrências $m_1 \ldots m_n$ de *M* seja idêntica a uma das ocorrências $f_1 \ldots f_n$ de F.

A oposição entre essas duas vertentes pode ser visualizada de modo mais claro com um exemplo simples. Considere as palavras que compõem esse texto. Naturalmente, todas elas se utilizam de determinados caracteres do nosso alfabeto para sua composição. Considere o termo "casa". Nesse termo, encontramos três caracteres distintos: "c", "a" e "s". Podemos dizer que essa palavra é constituídas por três *tipos* distintos de caracteres. Quando olhamos para os caracteres individuais, no entanto, vemos que não existem somente três deles, mas sim quatro. O que acontece é que um desses caracteres aparece duas vezes na palavra "casa". Assim, podemos dizer que há duas *ocorrências* de "a" na palavra "casa".

Imagine agora um caso em que escrevemos o termo "casa" em um caderno com uma caneta. Se olharmos para esse termo, veremos que os mesmos três tipos de caracteres se repetem. É possível notar, entretanto, que novas ocorrências dos caracteres acontecem nesse caso. Em outras palavras, temos aqui um caso em que os caracteres são realizados de maneira diferente (escritos por uma caneta) enquanto no caso anterior estavam impressos em uma folha sulfite. O que a teoria da identidade de ocorrências sustenta é que as diferentes ocorrências dos caracteres "c", "a" e "s" são idênticas na medida em que possuem o mesmo papel causal (ou funcional) na estrutura da qual fazem parte. Nesse sentido, não importa se escrevemos ou imprimimos a palavra "casa", em ambas situações ela terá o mesmo significado para um indivíduo que compreenda a língua portuguesa. O fato de ter sido escrita com tinta em caderno ou impressa em sulfite não altera o seu papel causal ou funcional. Similarmente, para a teoria da identidade de ocorrências, não importa se um estado mental do tipo *M* é realizado em material orgânico ou em uma estrutura de silício. O que importa é que as *ocorrências* de *M* em ambas as estruturas sejam causalmente ou funcionalmente idênticas.

Ao adotar a teoria de identidade de ocorrências, a proposta funcionalista parece ser sensível às objeções levantadas contra o behaviorismo e a teoria da identidade, uma vez que estados mentais são identificados com as relações causais e funcionais que exercem dentro de

um organismo. Isso não elimina, pelo menos em princípio, a hipótese de que animais localizados em patamares inferiores da escala evolutiva ou até mesmo sistemas não-orgânicos possam realizar um estado mental como uma dor, por exemplo.

A definição de estados mentais de acordo com o papel causal ou funcional que eles possuem no funcionamento de um determinado sistema parece, no entanto, aproximar o funcionalismo do behaviorismo, visto que um estado mental deve ser entendido como o estado que exerce uma relação causal entre *inputs* ambientais e *outputs* comportamentais. Mas além de assumir a importância do papel causal entre $inputs$ e $outputs$, o funcionalismo considera ainda a relação de um estado mental *M* com outros estados mentais que fazem parte de *S*. Tendo isso em vista, é possível notar que o funcionalismo se compromete com uma tese não-reducionista sobre a natureza da mente, visto que assume a importância do papel causal exercido por um estado mental no funcionamento de um organismo como um todo, e essa relação causal entre estados mentais não pode ser reduzida a nenhum meio realizador desses estados. Nesse sentido, o funcionalismo, conforme aponta Kim (1998), apresenta uma *concepção holista* da mente.

3.3.1 Objeções ao funcionalismo

Assim como no caso da teoria da identidade, trataremos aqui de duas dificuldades importante colocadas ao funcionalismo: (i) a objeção dos $qualia$ invertidos e (ii) a objeção do Quarto Chinês[7].

Os qualia invertidos

Uma das objeções mais conhecidas e que traz sérios problemas para uma abordagem funcionalista da mente consiste na objeção dos $qualia$ invertidos. De um modo geral, a objeção sustenta que dado que, para o funcionalismo, estados mentais são definidos de acordo com seu papel causal, parece não haver espaço nessa concepção para os aspectos fenomenais ou qualitativos da consciência.

[7] Ver Kim (1998) e Searle (2004) para discussões mais detalhadas desses problemas

Essa dificuldade é expressa por uma experiência de pensamento bastante conhecida na história da filosofia. O experimento de pensamento dos *qualia* invertidos, como se tem recentemente referido a esse cenário, propõe que imaginemos um mundo no qual o espectro de cores seja invertido, de modo que, se um sujeito *S* estiver olhando para um morango, ele não verá um objeto de cor vermelha, mas sim um objeto de cor verde. Nesse contexto, se adotarmos uma postura funcionalista em relação ao estudo da mente, uma consequência inevitável seria a de que a nossa perspectiva de estudo não será sensível à inversão dos *qualia*, uma vez que não poderíamos dizer se sistemas isomórficos em relação aos seus estados causais internos experimentam um mesmo *quale* referente ao vermelho ou ao verde, como no exemplo do morango. Isso ocorre porque os indivíduos do mundo em que o espectro se encontra invertido poderiam continuar dizendo que um morango é vermelho ainda que sua experiência subjetiva seja aquela que associamos à cor verde. Na verdade, o problema dos *qualia* invertidos é ainda mais expressivo se pensarmos que o modo que conhecemos e nos comunicamos sobre as cores dos objetos é, essencialmente, um modo ostensivo. Em outras palavras, termos linguísticos como "azul" e "verde" são definidos pelo apontamento dos objetos que são azuis ou verdes. Assim, é perfeitamente possível que um de nossos familiares ou amigos tenha nascido com o espectro de cores invertido, mas que se comuniquem normalmente dizendo que "o céu é azul" e a "a grama é verde", ainda que suas experiências subjetivas sejam diferentes das nossas.

Esse tipo de conclusão parece indicar que o funcionalismo, apesar de superar os problemas colocados às teorias que o precederam, ainda não é capaz de fornecer uma explicação efetiva da mente, uma vez que, segundo a objeção dos *qualia* invertidos, uma explicação funcionalista do mental deixaria de lado um fato importante do mundo, a saber, a possibilidade de haver uma diferenciação de $qualia$ em sistemas semelhantes no que diz respeito ao seu aspecto causal e funcional.

O Quarto Chinês

Em *Minds, Brains, and Programs* (1980), John Searle descreve seu famoso argumento do Quarto Chinês no qual pretende expor a falibilidade do programa de pesquisa que ele de-

nomina de “Inteligência Artificial Forte”. Searle propõe que pensemos em uma pessoa dentro de um quarto realizando a operação transitória causal entre os *inputs* ambientais, como frases em chinês recebidas por uma janela no quarto, e os *outputs* comportamentais enviados por uma pequena abertura na porta, como a emissão de frases em chinês coerentes com os *inputs*, guiada apenas pela mera operação sintática de símbolos e regras previamente estabelecidas. Nesse cenário, o indivíduo dentro do quarto dispõe de um conjunto de regras pré-estabelecidas que ditam a transformação dos termos em chinês provenientes em respostas (também em chinês) que sejam coerentes com os *inputs* iniciais.

O que Searle pretende mostrar com essa experiência de pensamento é que embora o comportamento do quarto pareça sugerir que seu funcionamento implica a compreensão do chinês, uma vez que a relação causal observada entre *inputs* e *outputs* é uma relação sintaticamente e semanticamente coerente, o próprio processo causal responsável por essa transição não compreende chinês. Isso ocorre porque os processos internos do quarto são apenas processos sintáticos e operacionais guiados por um conjunto de regras anteriormente estabelecidas. Nesse sentido, Searle afirma que a mera operação ou habilidade sintática não é condição *suficiente* para que o significado das emissões em chinês seja estabelecido.

É importante ressaltar que Searle está se referindo a uma tese específica do funcionalismo, segundo a qual a analogia entre um *software* e um *hardware* computacionais é colocada em paralelo com a analogia entre mente e cérebro. A mente, nessa concepção, seria um *software* “instalado” no *hardware* do cérebro. Embora nem todos os teóricos funcionalistas se comprometam com essa tese, a insuficiência de uma abordagem funcional para a compreensão dos aspectos semânticos da mente parece ser bastante preocupante. Isso ocorre porque o funcionalismo parece não poder explicar nem aspectos fenomenais e nem os aspectos semânticos da mente, o que o coloca em sérias dificuldades enquanto uma concepção plausível da mente.

1.2.4 Dualismo de propriedades

O dualismo de propriedades consiste na tentativa de situar os aspectos qualitativos dos estados mentais entre as propriedades fundamentais do mundo tais como massa e energia. Em outras palavras, as propriedades qualitativas dos estados mentais seriam fatos “brutos” do mundo no

sentido em que não podem ser reduzidas a termos mais fundamentais da física. Isso não quer dizer, no entanto, que não há uma relação de dependência das propriedades mentais com entidades físicas, como é o caso do cérebro. Nesse sentido, para o dualismo de propriedades o cérebro pode ser considerado o local onde a mente tem sua origem. A grande vantagem dessa posição teórica reside no fato de que ela reconhece a forte relação de dependência entre mente e cérebro sem pressupor a existência de uma substância distinta que seja o fundamento metafísico dos processos mentais. Outra importante vantagem do dualismo de propriedades reside no âmbito fenomenal dos estados mentais. Ao contrário das outras teorias aqui apresentadas, o dualismo de propriedades parece, pelo menos em princípio, ser sensível aos aspectos qualitativos dos estados mentais. Mas, diferentemente das teorias anteriores, o dualismo de propriedades se depara com problemas que resultam dessa tentativa de estabelecer propriedades mentais como propriedades básicas do mundo.

Isso pode ser visto de modo mais claro ao considerarmos formulações mais específicas do dualismo de propriedades. Essas formulações diferem, em sua essência, no modo em que concebem a relação das propriedades mentais com as propriedades físicas. Uma primeira forma de dualismo de propriedades estipula que as propriedades mentais são epifenômenos de propriedades físicas. Chamemos essa concepção teórica de *epifenomenalismo*. De acordo com o epifenomenalismo, o cérebro é responsável por dar origem à mente, mas, na medida em que ela é originada, ela passa a ter "vida própria" no sentido em que não possui poder causal sobre o cérebro. Em outras palavras, embora o cérebro seja responsável pela existência da mente, esta última não exerce nenhuma influência causal sobre o primeiro. Conceber o dualismo de propriedades desse modo permite evitar o problema da sobredeterminação causal mencionado anteriormente, sem, no entanto, abrir mão da existência dos estados mentais.

Um segundo modo de se conceber o dualismo de propriedades consiste em pensar as propriedades mentais como propriedades *emergentes* das propriedades físicas. Chamemos essa concepção de *emergentismo*. De acordo com o emergentismo, a organização de determinadas porções de matérias de uma forma específica e complexa seria capaz de dar origem a estados mentais, de modo que esses últimos, ao contrário do que diz o epifenomenalismo, não perderiam seus poderes causais. Um problema com o emergentismo é que não é tão claro como devemos

entender a noção de emergência. Em outros termos, no que ela difere da noção de redução? Além disso, apelar para a relação de emergência não parece assegurar aos estados mentais seu poder causal de modo não problemático. Para que possamos entender como a emergência evita a sobredeterminação causal, seria preciso que outras considerações fossem feitas sobre a natureza da mente e de sua relação com o mundo físico.

Embora o dualismo de propriedades possa ser formulado de outras maneiras, destacamos as duas acima em função do fato de que postular a existência de propriedades mentais em um nível básico requer uma explicação posterior de como elas se relacionam com o resto do mundo físico. Nesse sentido, o dualismo de propriedades dificilmente resolve todos os problemas apenas pela aceitação da existência da mente. Para que possa ser aceito como alternativa teórica séria, é preciso que termos como propriedades emergentes ou propriedades epifenomênicas sejam esclarecidos de modo rigoroso. Não tentarei realizar essa tarefa aqui, mas ao contrário, tentarei mostrar de modo breve como essas noções incorrem em dificuldades semelhantes às imputadas às teorias consideradas anteriormente.

Objeções ao dualismo de propriedades

Embora o dualismo de propriedades ofereça uma perspectiva interessante para concebermos uma teoria científica da mente no que diz respeito aos aspectos fenomenais, essa perspectiva positiva reflete dificuldades em outros âmbitos igualmente importantes no estudo da natureza da mente. A primeira dificuldade que o dualismo de propriedades enfrenta diz respeito às implicações contraintuitivas aduzidas pelo epifenomenalismo. Segundo essa objeção, parece ser bastante claro que estados mentais possuem uma relação causal com estados físicos. Mas se o epifenomenalismo for verdadeiro, uma consequência direta dessa concepção seria a de que quando desejo levantar meu braço, o meu movimento corporal, a saber, o movimento de levantar meu braço, não é causado pelo meu desejo, mas antes, é causado por meras interações elétricas que ocorrem em meu cérebro. Isso nos levaria a uma concepção de mente que nega a possibilidade de agirmos de acordo com aquilo que ponderamos em nossa mente, o que traz uma consequência tão pouco intuitiva quanto negar que nossos estados mentais possuem aspectos

qualitativos. Nesse sentido, enquanto capaz de salvar os aspectos qualitativos, o epifenomenalismo teria que abrir mão da eficácia causal da mente.

A segunda dificuldade referente ao dualismo de propriedades se apresenta ao emergentismo. Como vimos, o emergentismo, ao contrário do epifenomenalismo, tenta resguardar a eficácia causal dos estados mentais. Isso, no entanto, incorre em problemas como o da sobredeterminação causal em relação aos movimentos corporais. Se os estados mentais forem propriedades fundamentais do mundo, e, portanto, não-físicas, seguir-se-ia que o princípio do fechamento causal do universo físico seria violado, uma vez que algo não-físico exerceria influência causal em algo puramente físico. Nesse sentido, ainda que aceitemos o emergentismo na tentativa de superar as dificuldades do epifenomenalismo, teríamos que lidar com a indesejável consequência de que essa concepção implicaria a negação de um dos princípios mais básicos da física.

1.3 Conclusão

Como conclusão, uma última observação merece destaque. Ela se refere ao fato de as abordagens apresentadas até aqui serem capazes de superar dificuldades específicas, deixando de lado, no entanto, aspectos essenciais para o estudo da mente. Acredito que dentre as principais abordagens apresentadas, o funcionalismo se destaca como perspectiva teórica mais promissora. Como vimos, o funcionalismo também enfrenta dificuldades, mas ele parece superar algumas objeções importantes feitas ao behaviorismo e a teoria da identidade. No próximo capítulo, tratarei dessas questão com mais detalhe e tentarei mostrar porque o funcionalismo é preferível frente as outras alternativas. Essa discussão nos permitirá compreender melhor o problema epistêmico dos *qualia*.

Capítulo 2

Funcionalismo e o problema epistêmico dos *qualia*

2.1 Argumentos em favor do funcionalismo

Apresentei no capítulo anterior quatro das mais influentes teorias da mente que surgiram no último século. A discussão que apresentei foi de natureza geral, tendo como objetivo tornar explícitos alguns dos problemas que uma teoria científica da mente deve enfrentar. Em resumo, do que vimos até agora, parece haver três grandes problemas filosóficos que uma teoria da mente deve ser capaz de responder: (i) o *problema da causação mental*, isto é, como algo material pode ter influência causal em algo mental e vice-versa; (ii) o *problema da intencionalidade*, que questiona sobre a possibilidade de o cérebro, algo material, ter estados mentais que se *direcionam*, ou, mais ainda, que *representam* coisas do mundo físico; e (iii) o *problema dos qualia*: como é possível que algo material como o cérebro dê origem a estados mentais com aspectos qualitativos?

Como enfatizei anteriormente, o foco da minha discussão nessa dissertação será o terceiro problema. Deve estar claro, no entanto, que ao discutir esse problema não pretendo apresentar uma análise exaustiva do mesmo. Não tentarei, pelo menos não em um sentido direto, esboçar um cenário teórico que pretenda lidar primariamente com questões metafísicas ou ontológicas sobre os *qualia*. O meu objetivo é, ao contrário, oferecer uma *teoria epistêmica* dos

$qualia$ que lide com problemas epistemológicos levantados por eles. Nesse sentido, o meu objetivo geral consiste em mostrar que é possível superar os problemas epistemológicos levantados pelos $qualia$ dentro da ciência sem pressupor uma teoria ontológica bem definida sobre sua natureza.

Tendo esclarecido esses pontos, podemos agora prosseguir em nossa análise. Do que vimos até agora, parece ser plausível dizer que o funcionalismo como foi caracterizado aqui parece ser a melhor alternativa filosófica para se conceber sobre a natureza da mente. Isso ocorre porque, primeiro, o funcionalismo supera as objeções feitas ao behaviorismo e à teoria da identidade sem cair nos excessos metafísicos do dualismo de propriedades. Consideremos esse ponto mais calmamente. Vimos que o behaviorismo concebido como uma análise dos estados mentais em termos de comportamentos atuais ou disposicionais implica uma circularidade em sua análise. Em outras palavras, parece não haver um modo não-arbitrário de se estabelecer quando a análise de crenças e desejos de um sistema pode cessar. Se isso for verdade, então o behaviorismo nos leva, de modo necessário, a um regresso ao infinito.

O funcionalismo, tal como concebido por Putnam (1967, 1973) não enfrenta esse problema. Para vermos isso de modo mais claro, uma definição mais precisa do funcionalismo pode ser instrutiva. Essa definição pode ser dada nos seguintes termos:

O funcionalismo em filosofia da mente considera estados mentais como:

> (i) estados funcionais estabelecidos a partir da relação causal existente entre os $inputs$ que um organismo recebe do seu ambiente e os $outputs$ que ele emite como resposta a esses estímulos; e
>
> (ii) estados funcionais definidos a partir da relação que um estado mental *m* possui com o conjunto de outros estados mentais $m_1 \ldots m_n$ constituintes do organismo em questão.

Nessa perspectiva, o funcionalismo estipula que a determinação de um estado mental enquanto um estado funcional a partir de (i) e (ii) descrevem a própria natureza do mental. Essa assunção teórica constitui o ponto forte do funcionalismo em relação às teorias que o

precederam: isto é, os estados mentais, para o funcionalismo, são neutros quanto à natureza do substrato em que são realizados, não sendo dependentes, portanto, do substrato físico e de suas contingências. Os estados mentais são, nessa perspectiva, definidos de acordo com seu papel funcional exercido na economia de um organismo.

Aqui é importante entendermos no que exatamente essa vantagem teórica consiste. Para que isso se torne mais claro, além do fato de estados mentais serem neutros quanto a seu substrato, uma outra vantagem bastante importante consiste no fato de que eles são passíveis de uma *definição holística* em relação aos outros estados mentais aos quais estão relacionados. Comecemos pela análise do primeiro caso. Vimos que o behaviorismo e a teoria da identidade enfrentam uma dificuldade que chamamos de "chauvinismo especista". Essa dificuldade é expressa na medida em que nos damos conta de que uma análise dos estados mentais em termos de comportamento ou atividade cerebral é bastante limitada, uma vez que só permite atribuir mente para espécies de seres vivos que sejam consideravelmente semelhantes aos seres humanos. Assim, se defino um estado mental como a dor a partir de um conjunto de estados comportamentais, comprometo-me com a tese segundo a qual somente sistemas que podem realizar esses estados comportamentais na presença de determinados $inputs$ podem ter dor. Essa não parece ser, entretanto, uma conclusão plausível, visto que parece razoável imaginarmos criaturas advindas de um local muito longe do universo que tenham dor mas que não exibam nenhum dos estados comportamentais que associamos à definição de dor. Mais ainda, parece ser uma possibilidade em aberto que no futuro possamos construir computadores que sintam dor mas que não realizam os estados comportamentais que associamos aos humanos.

Quando tratamos da teoria da identidade, a objeção se coloca de modo similar. Se definirmos a dor como um estado cerebral como estimulações de fibras-C, essa definição implica a tese segundo a qual seres que não possuem uma estrutura cerebral similar ou idêntica à dos seres humanos não podem ter dor por definição. Podemos dizer que, no caso da teoria da identidade, a objeção parece ser ainda mais devastadora do que no caso do behaviorismo. Isso acontece porque o teórico da identidade deve negar a presença de estados mentais até mesmo para animais que se comportam de modo muito similar a nós quando estão na presença de $inputs$ relevantes. Se um extraterrestre qualquer, ao tocar uma chapa quente, estremecer e

gritar, tal como fazemos quando sentimos dor, o teórico da identidade não poderá atribuir a esse sistema o estado de dor, uma vez que suas estruturas internas diferem substancialmente da estrutura do cérebro humano.

Assim, uma vez que essas questões se tornam explícitas, uma definição funcional dos estados mentais parece apresentar um modo de evitá-las. Quando definimos estados mentais pelo papel funcional que exercem na relação entre certos $inputs$ e $outputs$, permanecemos neutros sobre a natureza do substrato material no qual esses estados são realizados. Considere o exemplo de um artefato comum, como é o caso de um garfo. Dado que um garfo realize sua função corretamente (auxiliar no manejo da comida), pouco importa se ele é feito de ferro ou de plástico. O que importa para a identificação daquele objeto como um garfo é a função que ele realiza, e não propriamente o material do qual ele é feito. Similarmente, o que faz de um estado mental uma dor não é o substrato material em que ele ocorre, mas sim a função que ele exerce na relação entre $inputs$ e $outputs$ do organismo do qual ele faz parte.

Torna-se claro, portanto, que o funcionalismo permite a múltipla realização dos estados mentais em um número amplo de espécies, desde que os aspectos (i) e (ii) descritos acima não sejam violados. Essa conclusão tem duas consequência de bastante importância: primeiro, não há mais uma limitação teórica para atribuirmos estados mentais aos animais que se encontram em degraus mais longínquos da história evolutiva, desde que eles tenham comportamentos semelhantes aos nossos na presença de $inputs$ relevantes. Mais ainda, não há uma limitação teórica para atribuirmos estados mentais a possíveis seres que não partilham da mesma estrutura material que nós, como é o caso de possíveis extraterrestres e robôs inteligentes. Esses dois pontos tornam bastante claro o modo em que o funcionalismo pretende superar as dificuldades enfrentadas pelo behaviorismo e pela teoria da identidade. Nesse sentido, podemos dizer que a tese da múltipla realização dos estados mentais consiste no ponto fundamental que diferencia o funcionalismo de alternativas anteriores.

Tendo discutido o primeiro caso, passemos agora ao segundo caso mencionado acima. Esse caso tem a ver com a definição holística dos estados mentais. Essa definição consiste em uma resposta ao problema da circularidade apresentado ao behaviorismo. De acordo com a objeção da circularidade, o behaviorismo não pode apresentar uma análise finita dos estados

mentais apenas em termos comportamentais. O funcionalismo pretende superar essa dificuldade pela introdução da tese (ii) em conjunção com (i). De acordo com (ii), estados mentais não são definidos somente pela relação funcional entre $inputs$ e $outputs$, mas também pela sua relação com outros estados mentais que fazem parte do sistema. A dor que tenho ao tocar uma chapa quente não é, nesse sentido, definida apenas pelos $inputs$ (danos na minha pele) e os $outputs$ (estremecimentos, gritos, etc.), mas também pela relação que esse estado mental tem com os outros estados mentais. Assim, além da relação funcional entre $inputs$ e $outputs$, a minha dor também é definida em relação às minhas crenças, desejos, emoções, etc. Esse tipo de definição situa os estados mentais num todo composto por outros estados mentais que estão mutuamente conectados, de modo que a definição da natureza de um estado mental nunca é dada em referência a estados individuais, mas sim no contexto da totalidade de estados mentais de um sistema. Podemos dizer, portanto, que o funcionalismo evita a objeção da circularidade porque ele permite que estados mentais sejam definidos em relação a outros estados mentais.

Como conclusão, parece seguro dizer que a múltipla realização e a definição holística dos estados mentais nos dão bons motivos para acreditar que o funcionalismo apresenta vantagens teóricas sobre teorias anteriores. Embora seja mais sofisticado nesses pontos, o funcionalismo ainda enfrenta difíceis objeções. Como vimos anteriormente, essas objeções se centram em dois tópicos principais: *intencionalidade* e *qualia*.

Como mencionei anteriormente, não tratarei do problema da intencionalidade neste trabalho. Uma teoria completa da relação entre mente-corpo depende intimamente da apresentação de respostas aos três principais problemas destacados acima (causação mental, intencionalidade e $qualia$). Para os propósitos desse trabalho, no entanto, terei como foco somente o problema dos $qualia$. Essa opção se dá não em função de uma predileção teórica que estabelece uma hierarquia entre esses problemas, mas sim pela necessidade prática de limitar o escopo do trabalho. Tendo isso em vista, na próxima seção apresentarei uma caracterização filosófica do problema dos $qualia$ e mostrarei como ele se relaciona com o funcionalismo de modo mais específico.

2.2 O problema ontológico e o problema epistêmico dos $qualia$

Existem pelo menos dois modos diferentes a partir dos quais podemos olhar para os problemas que os $qualia$ oferecem para o funcionalismo. Embora essa distinção não deixe de ser problemática, ela se faz útil para destacar os problemas que pretendo tratar neste trabalho. Nesse sentido, a distinção que apresento aqui deve ser entendida dentro desse contexto específico, e não como uma distinção absoluta. Desse modo, uma das principais preocupações que um estudo científico da mente tem que enfrentar é sobre a *ontologia* dos *qualia*. O problema pode ser colocado da seguinte maneira:

> **Problema Ontológico (PO)**: pode a ciência explicar como entidades de natureza qualitativa ($qualia$) têm sua origem e natureza explicada por entidades físicas mais básicas?

Em um trabalho bastante influente, David Chalmers (1996) denominou esse problema de o *problema difícil da consciência*. De acordo com Chalmers, uma explicação satisfatória da consciência deve mostrar como o cérebro (ou qualquer outro sistema físico) é capaz de originar e explicar a natureza da consciência[1]. Essa questão, como sustenta Chalmers, não parece ser uma questão trivial, uma vez que parece ser possível, pelo menos em princípio, conceber a consciência sem fazer referência ao mundo físico. O PO, nessa perspectiva, deve explicar em termos físicos como a consciência e todas suas características são possíveis no mundo natural.

O PO parece envolver questões que extrapolam aquilo que usualmente concebemos como objeto de análise filosófica, entrando diretamente no âmbito da investigação empírica. Isso não quer dizer, todavia, que o problema ontológico não pode ser, pelo menos parcialmente, estudado a partir de um ponto de vista filosófico. Podemos entender isso na medida em que atentamos para o fato de que o PO requer não somente uma análise de como estruturas físicas (como o cérebro) podem causar ou dar origem à consciência, mas também uma análise do que entendemos exatamente pelos termos "físico" e "consciência". Enquanto a investigação empírica empreendida pelas ciências cognitivas são essenciais para estabelecer os detalhes dessa relação,

[1] Considero aqui os termos "consciência" e "mente" como equivalentes. Embora tenhamos falado de mente até então, muitos debates na literatura, como é o caso de Chalmers (1996), fazem uso do termo "consciência" para se referir àquilo que venho chamando mais amplamente de mente. Assim, na ausência de considerações específicas, o leitor deve considerar ambos os termos como coextensivos no resto desta exposição.

não podemos eliminar de antemão a possibilidade de o PO derivar, pelo menos em parte, dos limites inerentes às categorias que utilizamos para tratar do problema em questão. É possível, como sugere Dennett (1988), que a nossa noção de $qualia$ esteja equivocada, ou, ainda, que a fundamentação filosófica da divisão entre mundo físico e mundo mental impeça uma abordagem conciliadora (ver SEARLE, 1992).

Embora o PO não seja o objeto de nossa discussão, essa breve caracterização serve como ponto de partida para a identificação da nossa problemática central. Além de oferecer difíceis desafios ontológicos para o estudo da mente, os $qualia$ também colocam problemas que são de natureza *epistêmica*. Antes de prosseguir, tomemos um momento para entender mais claramente quais são esses problemas. Para isso, talvez seja útil retomar a distinção entre *perspectiva de primeira pessoa* e *perspectiva de terceira pessoa*. Como vimos anteriormente, dizemos que algo é o caso sobre a mente e sobre o corpo a partir de diferentes perspectivas. No caso da mente, partimos de uma perspectiva de primeira pessoa para julgar sobre os conteúdos de nossos estados mentais. Usualmente, essa perspectiva envolve a capacidade de introspecção e é caracteristicamente marcada pela sua inefabilidade, isto é, pela impossibilidade de outros sujeitos terem acesso a esse conteúdo de modo direto. Como argumentara Descartes, a mente é a primeira coisa de que temos conhecimento claro e distinto. Isso quer dizer que não podemos estar enganados sobre o conteúdo de nossos estados mentais. O indivíduo a quem esses estados mentais pertencem tem, portanto, acesso privilegiado a esses últimos.

O acesso privilegiado que temos às nossas mentes se encontra em forte contraste com o acesso que temos com os objetos físicos do mundo. O computador que vejo em minha frente ao digitar este texto também pode ser visto por outros seres humanos e outros animais. Ao olharmos para ele, podemos descrever com precisão razoável suas propriedades. Mais ainda, o conteúdo dessa descrição é mutuamente compartilhado por todos que possam olhar e pensar sobre esse computador. Isso explicita o diferente acesso que temos do corpo em relação à mente, isto é, no caso do corpo, temos um acesso de terceira pessoa no qual múltiplos sujeitos podem validar um julgamento qualquer sobre a natureza de um objeto físico. Em outras palavras, isso quer dizer que objetos físicos podem ser vistos de diferentes pontos de vista sem que nenhum deles tenha o estatuto de privilégio epistêmico.

Neste ponto, aquilo que chamarei de problema epistêmico dos $qualia$ se torna mais evidente. Se mentes são acessíveis somente do ponto de vista de primeira pessoa, e se mentes são apenas "caixas pretas" que somente nós podemos olhar para dentro, então como poderemos avaliar as teorias científicas que estabelecem que a mente é o cérebro? Como poderei dizer, por exemplo, que meus amigos ou familiares possuem mentes se somente eles têm acesso a suas próprias mentes? De modo similar, como posso dizer que a minha experiência visual de um tomate maduro tem os mesmos aspectos qualitativos que a sua experiência visual desse mesmo tomate? Como vimos anteriormente, uma inversão do espectro de cores é bastante plausível, o que nos permite duvidar se temos os mesmos $qualia$ quando olhamos para um mesmo objeto físico.

Tendo isso em vista, podemos formular o *Problema Epistêmico* da seguinte forma:

> **Problema Epistêmico (PE)**: como podemos *justificar* a nossa *crença* de que temos experiências com os mesmos aspectos qualitativos (*qualia*) quando em contato com um objeto físico O nas mesmas condições perceptuais C?

Depois dessas considerações, parece inevitável concluir que se aceitarmos a distinção de acesso epistêmico que temos da mente e do mundo físico, então um estudo científico dos $qualia$ — e, portanto, um estudo científico da mente — não é possível uma vez que os $qualia$ não podem, por definição, ser objeto de conhecimento intersubjetivo.

Temos agora formulado de modo claro o problema que será objeto desse trabalho. No que se segue, tentarei mostrar que essa distinção epistemológica sustenta as dificuldades colocadas ao funcionalismo no que diz respeito aos $qualia$. Na próxima seção, discutirei isso em mais detalhes e argumentarei que o funcionalismo em sua concepção tradicional (versão desenvolvida por Putnam) não é capaz de responder a essas objeções de modo satisfatório. Essa discussão permitirá o estabelecimento dos fundamentos da minha proposta de solução do problema. De modo bastante resumido e antecipado, apresentarei uma formulação alternativa do funcionalismo que pretende responder satisfatoriamente o PE.

2.3 Problemas com o funcionalismo

O funcionalismo, tal como o descrevi aqui, é uma doutrina metafísica sobre a natureza da mente que pretende tornar possível o estudo científico desta última. Nesse sentido, podemos entender a versão do funcionalismo que descrevi aqui como uma versão ampla dessa tese. Embora os funcionalistas estejam comprometidos com as mesmas teses básicas (i e ii destacadas acima), é possível identificar diferentes assunções teóricas feitas por versões específicas do funcionalismo. No caso do funcionalismo de Putnam (1967, 1973), podemos dizer que uma de suas grandes motivações se situa na possibilidade de reproduzir qualquer função computável em uma máquina de Turing universal. Podemos chamar as versões do funcionalismo que se fundamentam nessa possibilidade de *funcionalismo de máquinas* (*machine functionalism*). Como se trata da formulação mais popular do funcionalismo, tratarei dela como mais detalhes nessa seção.

O funcionalismo de máquinas (FM) depende da possibilidade de uma máquina de Turing universal. Mas o que seria exatamente essa máquina? Uma máquina de Turing consiste numa máquina bastante simples constituída por uma fita contendo símbolos específicos que são escaneados por um detector programado a partir de um algoritmo específico. Assim, dado uma função $f(x)$, os valores iniciais verificados na fita ($inputs$) apontarão para um resultado de outra parte da fita ($outputs$) a partir da interpretação de $f(x)$. Nesse sentido, se o detector é tal que $f(x) = 2x + 1$, e x corresponde a um único segmento da fita, então se o detector se encontra em um momento t_1 no qual $x = 1$, no momento t_2 o detector se direcionará ao segmento da fita em que $x = 3$, uma vez que para $x = 1$, $f(x) = 2.1 + 1 = 3$.

Além de possuir valiosas aplicações práticas, uma máquina de Turing oferece recursos importantes para a reflexão filosófica acerca da natureza da mente. Isso pode ser visto de modo mais claro na idealização do modelo de uma máquina de Turing. Essa idealização, chamada de *máquina de Turing universal*, enfatiza a possibilidade lógica de qualquer função computável ser representada em uma máquina de Turing com um algorismo e uma fita adequados. Assim, se estivermos inclinados a conceber estados mentais como estados funcionais, então, pelo menos em princípio, a mente humana poderia ser descrita por uma máquina de Turing universal, ainda que essa não seja uma realidade prática.

Tendo identificado esse aspecto teórico mais geral do funcionalismo — isto é, sua concepção baseada na possibilidade de uma máquina de Turing — podemos agora explorar com mais precisão algumas das objeções às abordagens funcionalistas dos $qualia$. Para compreender isso de modo mais claro, podemos nos ater a duas objeções famosas feitas ao FM. Essas objeções são motivadas por uma distinção importante introduzida por Ned Block (1995). Block sugere a distinção entre dois modos específicos pelos quais podemos entender a afirmação de que um certo estado é "mental" ou "consciente". O primeiro caso consiste em dizer que um estado é consciente a partir da consideração do conteúdo informacional que transmite para o sistema ou organismo do qual ele faz parte e que determina os $outputs$ desse sistema em um determinado organismo. Block propõe que chamemos casos como esse de casos no qual há uma *consciência de acesso* (*access consciousness*). O segundo caso pelo qual podemos falar sobre um estado consciente é a partir da noção de *consciência fenomenal* (*phenomenal consciousness*), que está relacionada aos aspectos qualitativos dos estados mentais que discutimos anteriormente.

É importante notar que a distinção de Block sustenta uma divisão entre os aspectos funcionais ou cognitivos (consciência de acesso) e os aspectos fenomenais ou qualitativos (consciência fenomenal) dos estados mentais. Se essa distinção estiver correta, então parece ser possível imaginar casos em que um sistema S seja uma cópia funcional idêntica do cérebro humano mas que não tenha nenhum estado mental com $qualia$. De um modo mais preciso, é concebível a partir dessa distinção que haja um robô que diga algo como: "Vejo uma maçã vermelha" quando confrontado com uma maçã vermelha, mas que ainda assim não tenha um estado mental com o aspecto qualitativo da vermelhidão, isto é, o robô não sabe *como é* ter a experiência de ver algo vermelho.

Não é de se espantar, portanto, que a distinção de Block tenha motivado várias experiências de pensamento na filosofia da mente. Imagine o caso de seres humanos que veem o céu como amarelo, a grama como vermelha e uma maçã madura como verde. Esses indivíduos, embora tenham experiências qualitativas diferentes daquelas que temos, continuam a descrever o céu como "azul", a grama como "verde" e a maçã como "vermelha". Em outras palavras, essas pessoas tiveram seu espectro de cores invertido, mas continuam a usar os mesmos termos

que nós usamos para se referirem às suas experiências qualitativas. Nesse sentido, parece ser razoável dizer, pelo menos em princípio, que nessa situação os indivíduos que são funcionalmente idênticos podem ter experiências subjetivas e qualitativas distintas. Desse modo, uma definição funcional dos estados mentais não seria sensível aos aspectos qualitativos de nossas experiências.

Note que o problema surge porque uma definição funcional dos estados mentais não diz nada sobre seus aspectos qualitativos. Assim, quando alguém tem um estado mental funcionalmente idêntico ao estado de dor que tenho, não podemos estar completamente seguros de que essa pessoa tenha um estado mental com o mesmo aspecto qualitativo da minha dor. Isso ocorre porque não há uma relação lógica ou necessária entre os aspectos qualitativos e os aspectos funcionais de nossos estados mentais. Desse modo, a relação entre eles é uma relação contingente, o que gera um sério problema para validarmos o nosso discurso sobre esses casos, uma vez que nada pode garantir que o aspecto qualitativo que tenho em um estado de dor seja o mesmo aspecto que você tem quando em um estado funcionalmente idêntico. Em outras palavras, o fato de que os conteúdo de nossos estados mentais são somente acessíveis a partir da perspectiva de primeira pessoa não nos permite identificar a variação dos $qualia$ nesses casos. Podemos chamar essa objeção de objeção dos *qualia invertidos*.

Outro cenário possível que pode ser imaginado a partir dessa distinção é proposto por Block (1980). Block afirma que podemos imaginar uma situação na qual a população da China esteja organizada de tal modo que cada indivíduo recebe uma tarefa específica a ser realizada em um sistema composto por toda a população chinesa. Block nos pede também para supor que esse sistema seja uma cópia idêntica do cérebro humano, sendo cada chinês equivalente a um neurônio. Assumindo esses dois pontos de partida, parece razoável dizer que esse sistema se comportaria igualmente a um ser humano, mas não parece igualmente razoável supor que ele tenha consciência fenomenal. A partir disso, Block conclui que a consciência no sentido fenomenal não pode ser reproduzida pela mera cópia dos aspectos funcionais do cérebro humano. Se esse fosse o caso, teríamos que supor que o sistema composto pela população da China seja consciente assim como humanos o são. Frente a dificuldade de se aceitar essa tese, Block parece estar seguro de que ainda que determinada organização funcional seja reproduzida, os $qualia$

podem estar ausentes. Chamaremos essa objeção de objeção dos *qualia ausentes*.

Novamente, é importante observar que o problema é similar ao dos $qualia$ invertidos. Em outras palavras, a caracterização funcional dos estados mentais não fornece uma descrição exaustiva desses últimos. Como no caso dos $qualia$ invertidos, a objeção dos $qualia$ ausentes também apela para o caráter contingente da relação entre consciência e organização funcional. Assim, se aceitarmos essas premissas junto à distinção de Block (1995), chegamos inevitavelmente a um cenário no qual não podemos ter certeza da natureza dos estados mentais de outras pessoas. Isso inviabilizaria, por sua vez, o estudo científico da mente a partir de uma concepção funcionalista.

Passemos agora a um terceiro argumento que pretende mostrar fragilidades na definição funcional de $qualia$. Esse argumento é dado por Frank Jackson (1982). Jackson pede-nos para imaginar o caso de Mary, uma supercientista que vive dentro de um quarto preto e branco desde quando nasceu. Mary faz parte de um experimento científico do futuro, e por isso não pode sair de seu quarto. Além disso, os objetos do quarto de Mary foram organizados de tal modo que ela só os enxerga em preto e branco. Jackson diz que Mary é uma cientista brilhante e que ela sabe tudo o que se há para saber sobre o mundo físico. Note que Mary, ao saber tudo sobre o mundo físico, sabe de modo preciso o que acontece no cérebro de uma pessoa quando ela está olhando para um objeto vermelho. Em outras palavras, Mary tem uma descrição física completa dos eventos que ocorrem no mundo quando alguém tem a experiência subjetiva de algo vermelho. Imagine agora que Mary seja libertada de seu quarto e ao sair de lá se depare com uma rosa vermelha. A questão que Jackson nos coloca é a seguinte: teria Mary aprendido algo novo quando teve a experiência do vermelho? Teria Mary uma nova experiência quando olhou para a rosa? Se estivermos inclinados a dizer que não, então o funcionalismo e outras doutrinas fisicalistas enfrentam um problema, uma vez que Mary sabia tudo o que se podia saber sobre o cérebro humano e sua estrutura funcional antes de sair de seu quarto.

Por fim, há uma quarta dificuldade levantada por Thomas Nagel (1974) em seu artigo clássico "*What is it like to be a bat?*". Nesse texto, Nagel pede para imaginarmos um cenário em que saibamos precisamente tudo o que se há para saber sobre o sistema nervoso de um morcego. Sabemos que morcegos estão adaptados a ambientes de baixa luminosidade, e por esse

motivo, desenvolveram a habilidade de se localizar a partir do som que eles próprios emitem e a subsequente comparação da frequência do som emitido com aquele que é reverberado pela superfície dos objetos a sua volta[2]. Esse processo é conhecido como *ecolocalização*. O que Nagel pretende destacar com esse cenário é que, caso soubéssemos tudo o que se há para saber sobre o sistema nervoso dos morcegos, então deveríamos saber como é ter a experiência de nos localizar espacialmente através do som. O problema é que ainda que tenhamos uma ciência completa dos morcegos, essa tese não parece ser uma consequência direta desse cenário. Como diz Nagel, ainda assim não saberíamos como é ser um morcego (*what it is like to be a bat*). Assim, uma descrição funcional do sistema nervoso dos morcegos deixaria um aspecto importante de sua vida interna de lado, isto é, o aspecto referente ao *ponto de vista* dos morcegos frente ao mundo.

Para concluir essa parte, note que os problemas gerados por essas experiências de pensamento são consequências diretas dos diferentes modos de acesso epistemológico que temos de nossas mentes e dos objetos físicos. O fato de que os $qualia$ não podem ser explicados a partir da organização funcional de um sistema cria um hiato epistemológico entre o físico e o mental que parece ser insuperável. Isso acontece porque é impossível em princípio saber o que acontece no interior da mente de uma pessoa, isto é, de saber o conteúdo de seu fluxo de consciência (*stream of consciousness*). Isso faz do nosso conhecimento de suas mentes, e consequentemente de seus $qualia$, um conhecimento altamente incerto. Nessa perspectiva, o único modo de eu conhecer os seus $qualia$ consiste em um cenário no qual eu veja o mundo do seu ponto de vista, mas isso é impossível porque só posso ter acesso ao seu ponto de vista do mundo a partir do meu próprio ponto de vista. No final, teríamos um caso em que eu vejo o seu ponto de vista do mundo a partir do meu ponto de vista e não um caso no qual eu tenho acesso direto ao seu ponto de vista. Esse último cenário requer que eu me torne você, deixando de ser, desse modo, um sujeito epistêmico distinto de você.

Nessa mesma linha de raciocínio, podemos dizer que o único modo de saber como é ser um morcego consiste em ser um morcego. Mas para que isso seja possível, eu teria que perder meu estatuto epistêmico como ser humano para me tornar um morcego. No caso de

[2]Para uma discussão mais detalhada, ver Akins (1993).

Mary, o único modo de ela saber como é ter de fato uma experiência da cor vermelha é tendo essa experiência. Isso quer dizer que Mary só conhece a cor vermelha verdadeiramente quando ela tem uma experiência de primeira pessoa dessa cor. Por fim, no caso da população chinesa, não podemos dizer com certeza se o sistema possui uma mente ou não, visto que, embora isso seja bastante contraintuitivo, não podemos ter acesso direto ao que poderia ser a mente desse sistema. Assim, o melhor modo de nos certificar dessa questão é apelar para a nossa intuição, mas esse método é bastante problemático na medida em que muitas coisas que já pensamos ser contraintuitivas se tornaram verdadeiras em momentos posteriores.

Para concluir, os argumentos aqui apresentados têm como objetivo criar um contexto de dúvida sobre a possibilidade de uma abordagem funcionalista para o estudo da mente humana. Esse contexto de dúvida servirá como motivador da minha proposta, isto é, apresentar uma reformulação do funcionalismo que possa superar as dificuldades colocadas aqui.

2.4 Funcionalismo e teleologia

Vimos anteriormente que o funcionalismo caracterizado a partir da noção de máquina de Turing sofre uma série de objeções no que se refere aos aspectos qualitativos de nossos estados mentais. Um motivo que pode estar por trás dessas dificuldades é explorado por William Lycan (1981). Lycan argumenta que a noção de função usada pelo FM é uma noção estritamente matemática que interpreta a noção de função como um "mapeamento" de conjuntos (LYCAN, 1981, p. 27). Essa interpretação, para Lycan, é bastante limitada na medida em que restringe o poder explicativo do FM. Vejamos isso com mais detalhes.

Como explicitei acima, o FM afirma que um estado mental é definido pela relação que ele exerce entre $inputs$ e $outputs$. Um estado mental é, nesse sentido, um intermediário que mapeia (de acordo com sua definição funcional) a relação entre elementos de um conjunto I ($inputs$) com elementos de outro conjunto O ($outputs$). Esse tipo de relação funcional é típica das máquinas de Turing. Para Lycan, no entanto, essa definição deve ser descartada para o funcionalismo na filosofia da mente. Lycan acredita que uma nova compreensão da noção de função é necessária para reformular a noção clássica de funcionalismo:

> Considerarei, de modo literal, a ideia de que entidades mentais devem ser caracterizadas funcionalmente, mas não no sentido matemático como o faz o funcionalista de máquinas, e sim no sentido em que identificamos entidades mentais (como um primeiro passo em direção a uma concretude estrutural) em referência aos papéis [$roles$] que eles possuem em promover os objetivos e as estratégias dos sistemas em que eles acontecem (LYCAN, 1981, p. 27, trad. minha)

Essa diferença na caracterização da noção de função nos leva a uma nova formulação do funcionalismo que não se vincula à definição matemática de função, mas sim a uma nova concepção dessa noção em que a identidade funcional dos estados mentais é dada em referência aos papéis ($roles$) que eles possuem em promover os objetivos e as estratégias dos sistemas em que eles acontecem.

Essa nova concepção de função é conhecida na filosofia da biologia como a *concepção etiológica de função*. De acordo com a concepção etiológica, a função de um traço biológico ou de um artefato é dada a partir da análise de sua história seletiva. Assim, quando dizemos que a função do coração é a de bombear o sangue, a atribuição funcional não é feita meramente porque corações têm um papel intermediário de mapeamento entre $inputs$ e $outputs$, mas também porque bombear o sangue é o *motivo* pelo qual corações foram selecionados ao longo da evolução biológica. Nesse sentido, a função do coração no corpo humano é definida a partir do fato de que ele ajuda a promover os *objetivos* e as *estratégias* desses indivíduos. De um modo mais preciso, o corpo só pode funcionar — e, portanto, promover seus objetivos e estratégias — porque o coração bombeia o sangue para todos os outros órgãos.

Podemos dizer, no contexto dessa breve exposição, que a concepção etiológica tem um caráter *teleológico*. Em outras palavras, a função de um traço biológico ou de um artefato é definida a partir do *fim* para o qual esse traço ou esse artefato foi *selecionado* a realizar. Vale enfatizar que a atribuição de um fim a um traço biológico, contrariamente ao que observamos no caso de artefatos, não requer a existência de um agente intencional por trás desse processo, uma vez que o caráter de finalidade da análise biológica só é dado depois que analisamos a história evolutiva desse traço. Desse modo, ainda que atribuamos um aspecto teleológico a explicações funcionais biológicas, não precisamos nos comprometer com a existência de um agente inteligente realizando o processo de seleção.

Isso nos permite observar o importante papel que a seleção natural tem na determinação

das funções biológicas. Naturalmente, se assumirmos que o cérebro é um produto da seleção natural, e mais ainda, se assumirmos que a mente se origina no cérebro, uma investigação baseada nesses pressupostos pode se mostrar bastante frutífera. Para os meus propósitos aqui, a questão que se destaca é se o funcionalismo caracterizado a partir da concepção etiológica de função pode oferecer respostas satisfatórias aos desafios colocados ao FM na seção anterior.

Uma possível resposta para os problemas acima parece residir no aspecto normativo das funções etiológicas. Em outras palavras, embora saibamos que a função de um traço biológico X é determinada por sua história evolutiva, sabemos também que essa função não é alterada caso X não a realize em uma situação específica. De modo mais preciso, o fato de um coração particular não ser capaz de bombear o sangue não implica que sua função deixou de ser a de bombear sangue. Nesses casos, não dizemos que o coração perdeu sua função, mas sim que ele não está funcionando bem, ou, para usar um termo técnico, que ele está *mal-funcionando* (*malfunction*).

Trazendo essas noções para a discussão sobre o problema epistêmico dos $qualia$, é possível notar que uma noção de função que seja normativa parece ser elucidativa nos casos dos $qualia$ invertidos e dos $qualia$ ausentes. Se considerarmos que um sistema S tem uma estrutura funcional e uma história evolutiva similar a de um ser humano, então os seus $qualia$ não podem ser invertidos ou simplesmente estar ausentes se não houver uma diferença funcional no sistema. Isso ocorre porque a noção de função é *normativa*: isto é, existem critérios normativos que determinam quando estados mentais de diferentes tipos são ou serão o caso e esses critérios são determinados pela história evolutiva desses estados. Em outras palavras, se um estado mental do tipo m estiver associado à experiência de ver a cor vermelha e for realizado no cérebro humano, seguir-se-ia que esse estado está necessariamente conectado aos processos físicos ocorrentes no cérebro, uma vez que ele foi selecionado para esse fim (acompanhar determinado padrão de interação neuronal no cérebro). Similarmente, se houvesse uma diferença entre esses sistemas, essa diferença teria que ser objetivamente observável, seja no nível comportamental ou neuronal, uma vez que caso m não seja uma experiência da cor vermelha, m estaria mal-funcionando, e a própria noção de mal-funcionamento requer variações observáveis, caso contrário o conceito tornar-se-ia trivial.

Se essas considerações estiverem corretas, então o hiato explicativo entre consciência de acesso e consciência fenomenal parece ser reduzido, uma vez que a relação entre os dois passa a ser uma relação de necessidade. Para que haja diferença fenomenal, é preciso que haja também diferença funcional. O estatuto de relação necessária é adquirido somente na medida em que usamos uma noção de função que tenha um aspecto normativo. Nesse sentido, se a distinção de Block (1980) pode ser colocada em questão frente ao uso da noção de função etiológica para tratar da natureza dos estados mentais, então parece ser plausível admitir que elas podem lançar luz sobre as objeções apresentadas acima, uma vez que elas dependem da distinção feita por Block. Por isso, proponho que consideremos o que chamarei de *teleofuncionalismo* daqui em diante.

2.5 Teleofuncionalismo

Vimos nas seções anteriores que o funcionalismo como definido por Putnam não é capaz de lidar com as objeções colocadas pelos cenários hipotéticos apresentados. Tentei mostrar, na última seção, que esse problema se deve ao uso de uma concepção restritiva da noção de função. Argumentei que uma concepção mais ampla, a concepção etiológica, tem recursos teóricos para reavaliar a plausibilidade dos argumentos apresentados contra o funcionalismo. No contexto dessa crítica, uma nova concepção de funcionalismo pode ser formulada, concepção que podemos chamar de *funcionalismo teleológico* ou simplesmente *teleofuncionalismo*.

O teleofuncionalismo não é uma concepção nova do funcionalismo. Na verdade, ele se encontra formulado de diferentes modos em textos clássicos como Dennett (1991) e Lycan (1996). O meu objetivo aqui é, portanto, trabalhar essa noção tendo em vista os meus propósitos e, mais ainda, mostrar como ela pode resolver um problema importante na filosofia da mente. Nesse sentido, consideremos uma definição mais detalhada do teleofuncionalismo.

O teleofuncionalismo, tal como o concebo aqui, faz uso da concepção etiológica de função. Desse modo, o teleofuncionalismo sustenta que a definição de uma entidade funcional depende da sua história de seleção (natural ou não)[3]. Assim, se considerarmos o cérebro como produto da evolução, e se considerarmos a mente como relacionada ao cérebro, então uma

[3] Argumento de modo mais detalhado em favor dessa tese em Sant'Anna (2014).

investigação funcionalista baseada nesses princípios pode apresentar novas perspectivas para pensarmos alguns problemas na filosofia da mente. O argumento que sustenta essa empreitada pode ser esquematizado da seguinte forma:

(P1) O cérebro é um produto da seleção natural;

(P2) Existe uma relação entre mente e cérebro;

(P3) O funcionalismo é uma doutrina sobre a mente que merece consideração;

Portanto,

(C) Investigar a relação entre mente e cérebro a partir dos pressupostos do funcionalismo é uma empreitada válida.

Acredito que P1 e P2 são premissas pouco controversas e argumentar em favor delas implicaria um desvio considerável de nossa temática principal. Em relação a P2, devemos entendê-la como uma tese que afirma a relação entre mente e cérebro, ainda que a natureza dessa relação seja desconhecida. No que diz respeito a P3, acredito que ela se justifica em função da grande consideração, não somente por parte dos filósofos, mas também por parte dos cientistas da computação e psicólogos, que o funcionalismo tem recebido na literatura recente.

Tendo isso em vista, a minha proposta consiste no uso da concepção etiológica de função para avançar o debate sobre o funcionalismo na filosofia da mente. Cabe ressaltar que não argumentarei em favor de P1, P2 ou P3 e nem em favor de suas subteses. Com isso quero dizer que não tratarei aqui da questão se os $qualia$ podem ser ou não objetos de seleção natural. Essa é uma questão relativa ao problema ontológico, e por motivos práticos, não posso tratar dela nesse texto. Nesse sentido, um pressuposto que gostaria de esclarecer logo de início é o de que o tratamento da questão epistêmica dos $qualia$ se dará num âmbito neutro sobre a natureza metafísica dessas entidades. Para que a relação entre o problema epistêmico e o teleofuncionalismo possa ser entendida, basta que aceitemos P1, P2 e P3, ainda que não especifiquemos a natureza da relação explicitada em P2. Em outras palavras, o que chamarei posteriormente de *teoria epistêmica dos qualia* depende da tese minimalista de que há uma *correlação* entre mente e cérebro, tese que, como argumentei acima, acredito ser bastante razoável.

2.6 Colocando o problema

Temos agora o pano de fundo conceitual necessário para compreender o problema que tratarei aqui. Antes de seguir em frente, proponho uma reformulação do problema baseando-me em uma analogia com outro problema bastante conhecido na filosofia. David Hume enunciou um dos grandes problemas que ocupou (e ainda ocupa) a agenda dos filósofos modernos. Hume observou que o modo em que apreendemos as relações entre as coisas do mundo se dá a partir de uma sucessão de ideias em nossa mente. Essas ideias estão em um fluxo contínuo, algo parecido com o "fluxo de consciência" de William James. O que Hume notou é que nessa constante sucessão de ideias que resultam do contato perceptual com o mundo há algumas ideias que trazem algumas regularidades em sua apresentação. Um exemplo claro é que sempre que aproximo minha mão do fogo, sinto minha mão se aquecendo gradualmente.

No vocabulário do dia a dia, dizemos que o calor que sentimos ao aproximarmos nossas mãos do fogo é *causado* pelo fogo. Em uma análise mais detalhada, podemos identificar pelo menos dois momentos distintos nesse processo: primeiro, há o movimento da mão em direção ao fogo (E_1) e, em seguida, sentimos nossas mãos aquecer (E_2). De modo formal, dizemos que (i) há um evento E_1 que antecede a ocorrência de E_2, e mais ainda, (ii) que E_1 é a *causa* de E_2 ter ocorrido.

Esse tipo de afirmação parece, no entanto, extrapolar aquilo que podemos apreender diretamente pela sucessão desses eventos. Em outras palavras, quando dizemos que E_1 causa E_2, o modo mais intuitivo de se interpretar essa afirmação consiste em conceber uma regularidade que liga E_1 a E_2 e que E_2 acontece somente se E_1 for o caso. O problema reside justamente no fato de que não há nenhum fator que nos permite concluir que E_1 é a causa de E_2, e que para E_2 seja o caso, E_1 também precise ser o caso. Isso ocorre porque E_1 e E_2 não possuem nenhuma conexão intrínseca, isto é, não possuem nenhuma conexão tal que a partir da análise de E_1, podemos saber que E_2 será o caso. Sabemos isso simplesmente porque observamos haver uma regularidade na aparição de tais eventos, mas essa regularidade extrapola a relação imediata entre E_1 e E_2.

Para Hume, não podemos saber com absoluta certeza que quando E_1 for o caso, E_2 será *necessariamente* o caso. O aspecto necessário dessa relação é dado pelo hábito que criamos ao

observar essas ideias sucessivamente em quantidades crescentes de casos particulares. Isso quer dizer que estamos tão acostumados a observar que E_2 se segue a E_1 que toda vez que E_1 for o caso, tomamos como certo que E_2 também será o caso. Nesse sentido, a relação necessária pressuposta entre E_1 e E_2 não consiste numa relação necessária absoluta, mas sim em uma relação $quase$ necessária.

Note que o termo "quase" acima é mais do que suficiente para desinquietar um filósofo. Como mencionei anteriormente, o tipo de raciocínio que fundamenta essa inferência é um raciocínio indutivo (no qual a verdade da conclusão depende de modo probabilístico da verdade das premissas). Quando digo que minha mão se aquece quando a aproximo do fogo baseado meramente em minhas experiências passadas, não estou construindo meu discurso sobre uma base inteiramente confiável. Ao contrário, quando digo isso, tomo como certo o fato de que a regularidade que observei no mundo em momentos *passados* também será o caso no *presente*. Esse tipo de justificação, no entanto, é bastante fraca na medida em que consiste mais em uma esperança de que os eventos do passado se repetirão de modo regular do que propriamente na constatação de uma relação explicativa. Isso ocorre porque as premissas que tomo como ponto de partida (os eventos do passado) não implicam necessariamente a conclusão (eventos do presente). Assim, embora seja muito provável que minha mão se aquecerá quando a aproximar do fogo, não há garantia absoluta de que esse será o caso.

Essa dificuldade foi levada a sério por Immanuel Kant. Em sua *Crítica da Razão Pura* (1781), Kant procurou solucionar o problema colocado por Hume. Na *Introdução* da CRP, o autor funda aquilo que ele chama de *filosofia transcendental* que, de acordo com Kant, tem como objetivo responder a seguinte questão: como são possíveis juízos sintéticos *a priori*? Para responder essa questão, Kant distinguiu, em sua *Introdução*, dois tipos de juízos que usamos para descrever as coisas: (i) os *juízos analíticos*; e (ii) os *juízos sintéticos*. De um modo geral, juízos analíticos são aqueles em que o predicado está contido na própria definição do sujeito. Assim, em um juízo analítico como "Todo solteiro é pessoa não-casada", notamos que o predicado "pessoa não-casada" está contido na definição do sujeito "solteiro". A relação entre sujeito e predicado é, desse modo, analítica, uma vez que um juízo analítico não expande nosso conhecimento, mas somente torna mais claro determinada noção a partir de sua análise. Como

há uma relação lógica intrínseca entre sujeito e predicado, dizemos que esses juízos são válidos *a priori*, no sentido em que não precisam recorrer à experiência.

No caso dos juízos sintéticos, a relação entre sujeito e objeto não é uma relação intrínseca. Quando digo que "O meu carro é branco", uso essa proposição para expressar algo sobre um estado de coisas do mundo. Note que nesses casos o predicado não está contido no sujeito. Quando digo que tenho um carro, nada me diz qual é a cor desse carro. A relação entre o carro e sua cor não é uma relação dada no nível conceitual (o conceito de carro não implica o conceito de branco), o que exige que extrapolemos o âmbito da proposição para nos certificarmos se o meu carro é realmente branco. Nesse caso, dizemos que a experiência se torna juíza para estabelecer a conexão entre sujeito e predicado: temos que ir à garagem e ver se o carro é branco ou não.

Nesse ponto, podemos entender o que Kant queria dizer por *juízo sintético a priori*. Em outras palavras, Kant acreditava existir um terceiro tipo de juízo, aqueles juízos que expressam verdades necessárias mas que não são analíticos. Isso quer dizer que existem juízos que são necessários mas que a relação entre sujeito e predicado não é uma relação lógica intrínseca (o sujeito não contém o predicado). Um caso típico dessas expressões, de acordo com Kant, são os juízos da matemática e juízos causais, como "As mãos aquecem quando se aproximam do fogo". Em resumo, juízos sintéticos *a priori* são juízos que expressam relações necessárias que não são extraídas da experiência. Explicar a possibilidade desses juízos é de suma importância, visto que, como vemos a partir de Hume, não podemos inferir relações causais a partir da experiência.

Mas como essa discussão pode auxiliar no problema epistêmico dos $qualia$? Acredito que a distinção feita por Kant pode ser bastante instrutiva para os nossos propósitos. Considere novamente o problema difícil de Chalmers (1996). Chalmers sustenta que podemos conceber um mundo que é fisicamente idêntico ao nosso mundo, mas que nossas cópias físicas nesse mundo simplesmente não tenham consciência. Esses "doppelgangers" são conhecidos como zumbis filosóficos. O zumbi filosófico é uma entidade física idêntica ao ser humano e que age como nós, mas simplesmente não possui consciência. A sua vida interna é uma completa escuridão. De acordo com Chalmers, embora zumbis possam não ser nomologicamente o caso,

o mero fato de tal mundo ser logicamente concebível nos permite observar que a consciência é, em uma dimensão lógica, algo distinto do mundo físico. Isso não significa que a consciência pode existir sem um substrato físico, mas sim que o conceito de consciência não está contido no conceito de físico.

Não pretendo discutir se distinções lógicas implicam distinções ontológicas. Tal questão pode ser deixada de lado para os nossos propósitos daqui em diante. O que essa distinção mostra, no entanto, é como o problema epistêmico dos $qualia$ se relaciona à distinção apresentada por Kant. Se Chalmers estiver correto, então uma definição funcional da consciência não pode estabelecer uma relação necessária entre aspectos físicos ou funcionais e aspectos qualitativos dos estados mentais. Como enfatizei anteriormente, isso não significa que essa distinção possa ser sustentada em um nível nomológico. Pode ser o caso que as leis do nosso mundo sejam tais que dado uma estrutura funcional idêntica a do cérebro humano, a consciência necessariamente surgirá. Se esse é o caso ou não, não é algo que podemos dizer *a priori*. Temos que especificar quais são essas leis do mundo que permitem que a consciência seja uma consequência necessária da organização funcional ainda que essa relação não seja lógica. É nesse sentido, portanto, que afirmo que a teoria que desenvolverei aqui não se relaciona com os problemas ontológicos dos $qualia$, mas sim com problemas epistemológicos.

Desse modo, podemos dizer que o problema epistemológico dos $qualia$ tem a mesma estrutura lógica do problema relativo à possibilidade dos juízos sintéticos *a priori*: isto é, como podemos validar nossa crença de que outras pessoas possuem mente, ou, mais especificamente, como posso validar minha crença de que outras pessoas possuem os mesmos $qualia$ que possuo? Ou, ainda, como podemos justificar a nossa crença de que a relação entre identidade funcional e identidade mental possa ser necessária sem que essa relação seja analítica?

Como espero ter mostrado, outras respostas a essa questão falharam na medida em que se baseavam em um raciocínio de natureza indutiva em que a partir da observação de regularidades entre estados mentais e estados físicos, concluímos que outras pessoas também possuem estados mentais porque as mesmas regularidades se apresentam nos seus casos específicos. Não é de se surpreender, portanto, que muitos filósofos não se contentaram com essa conclusão, apresentando diversos cenários que pretendem mostrar sua fragilidade. Para concluir, minha

proposta nos próximos capítulos será guiada pela seguinte questão: como podemos justificar nossa crença de que organização funcional e $qualia$ (uma conjunção sintética) pode apresentar uma relação necessária sem ser analítica? Enfrentemos agora esse problema.

Capítulo 3

Funções etiológicas e funcionalismo

O último capítulo terminou com uma proposta de associar o funcionalismo na filosofia da mente com a concepção etiológica de função. Neste capítulo, tenho dois objetivos principais: primeiro, discutir sistematicamente essa noção e destacar alguns aspectos filosóficos associados a ela. Em particular, apresentarei aqui a noção de explicação teleofuncional que servirá de base para a teoria epistêmica dos $qualia$. Ao final dessa discussão, defendo a ideia de que se dois itens possuem uma mesma explicação teleofuncional, então ambos possuem a mesma função. O segundo objetivo consiste em relacionar, também de modo sistemático, a noção de função etiológica com o funcionalismo na filosofia da mente. A partir dessa discussão, apresentarei a minha definição de teleofuncionalismo que servirá de base para a formulação da teoria epistêmica dos $qualia$.

Tendo dito isso, comecemos por discutir a noção de funções etiológicas. O nosso vocabulário corriqueiro tem pelo menos dois modos importantes pelos quais podemos falar da função dos objetos. Itens como garfos e óculos foram designados para certos propósitos que, na maioria dos casos, têm o objetivo de nos ajudar a realizar determinadas tarefas mais precisamente. De um modo mais específico, podemos dizer que uma das tarefas associadas aos garfos é a de nos ajudar a pegar a comida em nossos pratos, ou, para usar outro exemplo, que a tarefa associada aos óculos é a de ajudar pessoas com problemas de visão. Essas atribuições funcionais podem ser expressas a partir de sentenças do tipo "A função de A é fazer/realizar B", na qual *A* corresponde a um objeto (um garfo, por exemplo) e *B* à tarefa ou processo que este objeto deve realizar (ajudar-nos a pegar a comida do prato).

É importante ressaltar que *A* não precisa ser um objeto no sentido usual do termo, isto é, objetos físicos como garfos ou óculos. *A* pode ser, por outro lado, um processo biológico como a fotossíntese ou a digestão. Embora difiram nesse aspecto, o sentido pelo qual falamos de funções em ambos os casos é similar: podemos dizer, por exemplo, que o propósito da fotossíntese é providenciar glicose para a planta, ou, de um modo mais específico, providenciar energia para a sobrevivência da planta. Um ponto interessante em relação a esses objetos e processos é que eles não precisam realizar as tarefas que foram associadas a eles, e, ainda assim, são considerados como se tivessem a função de realizar essas tarefas. Isso é mais claro no caso da biologia. Considere o caso da fotossíntese. Por alguma razão desconhecida, uma planta pode não conseguir realizar o processo completo de fotossíntese (por exemplo, ela não pode reduzir NADP para NADPH) porque algumas de suas partes não estão funcionando apropriadamente. Nesse caso, não dizemos que as células responsáveis por realizar a fotossíntese perderam sua função, mas sim que elas não foram capazes de realizar essa função. O ponto importante para retermos dessa discussão é que nesses casos de explicações funcionais parece haver a atribuição de um *fim* a esses processos, isto é, as funções parecem ser entendidas como determinações de um *télos* ou da *finalidade* desses objetos e processos.

Outros casos de atribuição funcional que encontramos no nosso dia a dia são casos em que consideramos as capacidades causais atuais ou disposicionais de um objeto. Considere novamente o caso dos garfos. Podemos, por exemplo, utilizar garfos para marcar as páginas de um livro. Embora essa não seja uma situação usual, parece ser perfeitamente concebível um cenário no qual as pessoas estejam acostumadas a usar garfos como marcadores de página. Nesse caso, embora garfos não tenham sido produzidos para servirem como marcadores de páginas, podemos dizer que garfos funcionam como se fossem marcadores de páginas. Um aspecto importante a se notar nesses casos é que as atribuições funcionais empregadas não fazem nenhuma referência aos fatos do passado do objeto ao qual se refere. Nesse sentido, as atribuições funcionais são feitas a partir de um *contexto*, isto é, dado um mundo no qual as pessoas usam garfos como marcadores de páginas, podemos dizer que a função dos garfos é a de marcar páginas nesse mundo. Fora desse contexto, no entanto, essa afirmação soaria um pouco estranha ou até mesmo sem sentido. Assim, podemos dizer que nesses casos as

atribuições funcionais podem ser descritas por sentenças como "A função de A é fazer/realizar B" associadas a uma referência ao contexto em que a sentença é emitida.

Essa breve discussão de algumas de nossas intuições acerca de atribuições funcionais servirá como base para as discussões que virão nas próximas seções. Na próxima seção em particular, apresentarei as formulações teóricas desses dois tipos de explicações funcionais dentro da filosofia da ciência. Essa discussão permitirá uma caracterização filosófica mais precisa da noção de *explicação teleofuncional*.

3.1 O que são explicações teleofuncionais?

As explicações teleofuncionais são muito utilizadas na biologia, principalmente dentro da área específica da biologia evolutiva. Como o próprio termo sugere, explicações teleofuncionais são explicações tanto *teleológicas* quanto *funcionais*. Para um leitor familiarizado com a história da filosofia e com a história da ciência, parece haver aqui uma tensão iminente. Em outras palavras, como é possível pensar em explicações científicas que tenham um aspecto teleológico depois de Darwin? Essa é, de fato, uma questão muito importante e que tem sido debatida extensivamente pelos filósofos da biologia. Não poderemos, no entanto, entrar nas minúcias desse debate, visto que extrapolaria os propósitos dessa dissertação. Podemos, entretanto, tentar entender a relação entre explicações funcionais e explicações teleológicas, o que será de extrema importância para compreendermos a noção de explicações teleofuncionais. Passemos, portanto, a essa discussão.

A década de 1970 foi muito importante para as discussões acerca do estatuto das explicações funcionais na filosofia da ciência. Podemos destacar duas noções particularmente interessantes que foram desenvolvidas nesse período. A primeira está associada ao trabalho de Larry Wright (1973). De acordo com Wright, as explicações funcionais, tanto na biologia quanto no caso de artefatos, são caracterizadas pela referência que elas fazem à história do traço biológico ou do artefato que é objeto de explicação. Como o próprio autor explicita, dizer que a função de *A* é realizar *B* quer dizer que:

(a) A existe porque realiza B;

(b) B é uma consequência ou resultado da existência de A (WRIGHT, 1973, p.

161)

Podemos notar, nesse caso, que uma explicação funcional que segue os preceitos de Wright explica tanto o porquê de *A* existir quanto o porquê de *A* realizar a função que realiza atualmente. Para entendermos isso de um modo mais claro, podemos recorrer ao caso clássico de uma explicação da função dos corações: corações existem porque bombeiam sangue (condição a); e o bombeamento de sangue é uma consequência da existência dos corações (condição b). A concepção de função que se origina no trabalho de Wright ficou conhecida como *concepção etiológica de funções*[1].

A segunda noção de função que surge na década de 1970 está associada ao trabalho de Robert Cummins (1975). Contrapondo Wright (1973), Cummins não acredita que explicações funcionais precisam fazer referência a fatos históricos relativos ao item a ser explicado. Para Cummins, uma explicação funcional deve ser restrita somente aos fatos atuais sobre o item em questão. Nesse sentido, utilizando-nos do caso dos corações novamente, pouco importa para um teórico que adota a concepção de Cummins se corações bombearam sangue no passado. O que importa, por outro lado, é que corações bombeiam sangue no contexto atual, e uma explicação funcional teria que explicar, em termos causais, como é possível que corações possam bombear o sangue. Uma forma de explicar como os corações podem bombear o sangue é empregando o que Cummins chama de *análise funcional*. As análises funcionais são muito comuns na inteligência artificial e na psicologia, sendo elas realizadas da seguinte maneira: quando queremos analisar a função de um sistema *S*, podemos decompor a tarefa *Y* realizada por *S* no trabalho das subpartes ($s_1, \ldots, s_n$) de *S* de tal modo que a tarefa complexa de realizar *Y* pode ser explicada pela operação de partes menores e menos complexas.

> A produção [Cummins se refere aqui às linhas de montagem] aqui é quebrada em tarefas distintas. Cada ponto na linha é responsável por uma determinada tarefa e é a função dos trabalhadores ou das máquinas que a tarefa seja realizada naquele ponto. Se a linha tem a capacidade de produzir o produto, ela tem essa capacidade em virtude do fato de que os trabalhadores ou as máquinas realizam as tarefas às quais eles foram designados, e em virtude do fato de que quando estas tarefas são realizadas organizadamente de um determinado modo — de acordo com um certo programa — o produto final aparece como resultado. Aqui nós podemos explicar a capacidade da linha de produzir o produto

[1]Para mais sobre o assunto, ver Neander (1991a), Neander (1991b), Kitcher (1993), Griffiths (1993), Millikan (1989a, 1989b, 1999, 2002), Buller (1998), Godfrey-Smith (1993) e Sant'Anna (2014).

> — i.e., explicar como a linha produz tal produto — apelando para certas capacidades dos trabalhadores ou das máquinas e à organização destas capacidades em uma linha de produção. (CUMMINS, 1975, p. 760)

Para se referir a essa concepção de função, isto é, às funções resultantes da análise funcional, os teóricos comumente se utilizam dos termos *funções de papel causal* ou *funções de Cummins*.

Muitos autores têm discutido extensivamente se é possível apresentar uma concepção de função que concilie esses dois projetos aparentemente distintos[2]. Para os nossos propósitos, no entanto, tal discussão pode ser deixada de lado. O meu interesse nesse momento é focar na primeira noção de função, visto que ela é central para as explicações teleofuncionais. Tentarei, na próxima seção, mostrar como a concepção etiológica de função está associada à noção de teleologia, o que será essencial para os nossos propósitos.

3.2 Funções etiológicas, teleologia e explicações teleofuncionais

Um importante aspecto das funções etiológicas é que elas possuem um aspecto teleológico[3]. Mas o que exatamente isso quer dizer? Para entendermos isso mais claramente, precisamos, primeiramente, entender o que se quer dizer por *teleologia* e como essa noção se relaciona às *funções etiológicas*. No contexto dessa discussão, podemos entender o aspecto teleológico das funções etiológicas como o estabelecimento de um *fim* ou um *télos* para o objeto de explicação. Assim, no exemplo utilizado acima, o fim ou o *télos* dos corações é justamente o de circular o sangue. Para usar a terminologia de Wright (1973), corações só existem *porque* são capazes de circular o sangue. Esse é o motivo pelo qual a seleção natural "selecionou" o tipo (*type*) corações.

Note que esse aspecto teleológico das funções etiológicas está diretamente associado à condição (a) estabelecida por Wright (1973). Dizer que corações existem porque possuem a função de bombear o sangue quer dizer que eles existem para um fim ou *télos* específico, a saber, bombear o sangue no organismo de alguns animais de tal modo que seu organismo possa

[2] Ver especialmente Kitcher (1993), Griffiths (1993), Godfrey-Smith (1993), Buller (1998), Millikan (1999, 2002) e Sant'Anna (2014).

[3] Ver Neander (1991a).

funcionar corretamente. Temos aqui, portanto, o ponto que nos permite relacionar a noção de teleologia enquanto presença de um fim ou *télos* à noção de função etiológica. São essas explicações funcionais etiológicas com um aspecto teleológico que chamarei de *explicações teleofuncionais*.

Agora que temos em mão o aparato teórico básico para entendermos o que são explicações teleofuncionais, podemos nos ater a uma análise um pouco mais detalhada do aspecto *lógico* dessas explicações. Para isso, será de grande ajuda nos atentarmos ao trabalho de William Wimsatt (1972). Embora Wimsatt não utilize o termo "explicações teleofuncionais", o seu trabalho explicita os principais aspectos lógicos desta última. Ele nos apresenta em seu artigo seis variáveis que têm como objetivo captar de modo detalhado todos os aspectos lógicos de uma explicação teleofuncional. Essas variáveis são: *item (i), sistema (S), ambiente (A), propósito (P), comportamento (C) e teoria (T)*.

Comecemos pela análise do *item*. Na verdade, não há muito para se dizer sobre essa variável. O item é o objeto que escolhemos para ser o objeto de uma explicação funcional. Wimsatt (1972, p. 19) ressalta que uma função é sempre função de algo, o que exige de uma explicação funcional um objeto de explicação. Para Wimsatt, existem duas classes centrais de objetos ou itens: (a) os objetos físicos ordinários; e (b) os comportamentos. O último caso está associado às explicações da psicologia evolutiva, que analisam um determinado comportamento atual a partir da presença desse mesmo comportamento em indivíduos ancestrais de uma determinada espécie.

Dizer meramente que a função de *A* é realizar *B* é uma explicação muito geral que não está livre de ambiguidades. Para exemplificar isso, Wimsatt (1972, p. 19) menciona o caso dos capilares periféricos nos mamíferos superiores. Uma das funções dos capilares periféricos é a de possibilitar a troca de elementos nutritivos e eliminar aqueles elementos que não são mais usados no sistema circulatório e nas células. Essa não é, entretanto, a única função que os capilares periféricos têm nos mamíferos superiores. Eles são responsáveis também pela regulação da temperatura nas áreas periféricas. Temos aqui um caso no qual precisamos decidir qual função será objeto do nosso estudo.

A partir desse exemplo, Wimsatt diz que explicações teleofuncionais precisam ser qua-

lificadas de um modo mais preciso. Para que isso seja possível, ele argumenta que é preciso especificar a função que atribuímos a *i* em relação a um *sistema (S)*. Nesse sentido, se os nossos interesses explanatórios estiverem direcionados à questão sobre como a troca de elementos se dá nos mamíferos superiores, então podemos considerar o primeiro caso de explicação como uma explicação genuína. Similarmente, se estivermos interessados em saber como determinadas áreas do corpo dos mamíferos superiores regulam a temperatura, então podemos aduzir ao segundo caso de explicação teleofuncional.

Essas qualificações são, sem dúvidas, muito esclarecedoras, mas elas ainda não são suficientes para eliminar toda a ambiguidade das explicações teleofuncionais. Podemos atribuir a um artefato ou traço biológico diferentes funções de acordo com o *contexto* no qual eles estão inseridos. Para entendermos isso, considere o caso dos garfos que discutimos no começo desse capítulo. Garfos são usualmente utilizados por nós para pegar a comida no prato, mas é perfeitamente plausível pensarmos em um caso no qual garfos possam ser usados como marcadores de páginas em um livro. Dado a caracterização de um contexto específico, não parece ser estranho dizer que a função dos garfos é a de marcar páginas. Isso mostra que a função de um item *i* depende do *ambiente* no que ele está inserido. Em outras palavras, especificar o sistema (S) em que *i* é realizado não é suficiente para evitar ambiguidades como a presente nesse caso. Precisamos mencionar, além disso, o ambiente em que *i* é realizado em *S*. Temos, aqui, a terceira variável de Wimsatt (1972), isto é, o *ambiente (A)*.

Neste ponto de nossa discussão, não seria estranho dizermos que esses três critérios ainda não são suficientes para evitar todas as ambiguidades. Se nos restringirmos apenas às três variáveis até aqui mencionadas, corremos o risco de tomar meros efeitos por funções genuínas. Isso fica mais claro no caso do nariz humano. A função do nariz humano claramente não é a de sustentar os óculos, embora possamos construir uma explicação funcional dessa relação nos baseando nas variáveis delineadas até aqui. Há aqui, portanto, um caso no qual podemos explicitar todas as três variáveis sem de fato capturar a função de um item (no caso, o nariz humano). Para resolver esse problema, é preciso que apelemos para uma nova especificação de modo que as diferenças entre funções e meros efeitos sejam captadas. A quarta variável que Wimsatt nos apresenta e que pretende resolver esse problema é o *propósito (P)*. Essa variável

faz referência explícita aos processos de seleção por trás do item *i*, o que a torna essencial para as explicações teleofuncionais. Em outras palavras, é por causa da atribuição de propósitos a *i* que podemos distinguir funções de meros efeitos. Quando dizemos que o propósito do nariz é do permitir o funcionamento adequado de determinados processos biológicos, excluímos deste grupo a mera condição de ser o suporte para os óculos.

Não surpreendentemente, *P* não é suficiente para eliminar todas as ambiguidades possíveis. Isso leva Wimsatt (1972) a mencionar uma quinta variável: o *comportamento (C)*. Para entendermos a importância dessa variável, considere o caso de um ar condicionado quente e frio. O propósito que os designers desse aparelho deram a ele é o de permitir ao usuário controlar a temperatura em um certo ambiente. A função do ar condicionado, no entanto, é tanto a de deixar o ar mais frio quanto a de aquecê-lo. Uma possível solução aqui poderia ser apelar para a variável *A*. Poderíamos dizer, por exemplo, que a função do aparelho é a de aumentar a temperatura quando estiver frio e diminuir a temperatura quando estiver calor. Essa consideração, no entanto, não resolve o problema, visto que alguém poderia, ainda que em um dia frio, querer diminuir a temperatura ainda mais. Para evitar essas dificuldades, Wimsatt (1972, p. 26) sugere que sejam incluídas considerações sobre o comportamento de *i*. De um modo mais específico, poderíamos qualificar a explicação funcional do ar condicionado dizendo que a sua função é a de aumentar a temperatura quando ele se comporta desse modo e o mesmo se aplicando no caso oposto. Nesse sentido, a atribuição funcional fica restrita ao contexto particular no qual empregamos nossa análise.

Finalmente, a última variável que Wimsatt (1972, p. 28) apresenta é aquela que especifica teorias científicas mais gerais que servem de *background* para as explicações funcionais. Para citar um exemplo, Wimsatt pede-nos para imaginar um caso no qual descobríssemos um organismo que tivesse sua estrutura tão diferente das que conhecemos atualmente, de tal modo que o controle da operação de seus vários órgãos não se desse por algo como um sistema nervoso. As operações desses órgãos se dariam, ao contrário, pelo som que eles emitem. Assim, supondo que essa criatura tenha um coração, a operação desse órgão iria emitir certos ruídos que guiariam os ruídos dos outros órgãos.

No caso acima, parece ser possível atribuir ao coração destas criaturas a função de pro-

duzir sons. De acordo com Wimsatt, no entanto, é muito improvável que esse seja o caso. Para sustentar essa afirmação, Wimsatt diz que uma explicação funcional deve estar de acordo com certas suposições de fundo (*background assumptions*), suposições que ele classifica como as "teorias causais" que fundamentam nossa investigação. Para entendermos isso, considere que a criatura que mencionamos acima seja um animal de sangue quente. Para que seus órgãos possam funcionar de modo adequado, seria preciso haver um modo seguro através do qual o som pudesse chegar aos outros órgãos. Dado nosso conhecimento das leis mais gerais da física, no entanto, sabemos que seria implausível dizer que o sangue seja o condutor de informações sonoras, visto que ele não é um bom condutor do som. Desse modo, ainda que o sangue seja um forte candidato para a circulação da informação sonora, uma vez que circula por todo o corpo do animal, essa possibilidade é excluída de antemão por teorias mais gerais da física que tornam essa hipótese muito pouco provável. Se, no entanto, nosso mundo diferisse de tal modo que o sangue pudesse se tornar um condutor eficiente do som, então essa hipótese seria uma hipótese muito forte. Por fim, o ponto geral dessa última variável parece ser o de evitar conflitos entre explicações funcionais e teorias mais bem estabelecidas na ciência. Uma hipótese pode ser muito atrativa e estar em conformidade com todas as nossas intuições, mas se ela estiver em conflito com alguma teoria geral mais bem estabelecida, então ela não deve ser uma candidata real para uma explicação funcional.

Tendo explicitado todas as variáveis de uma explicação teleofuncional, podemos agora estabelecer um panorama geral da nossa discussão. Primeiramente, vimos que as variáveis de uma explicação teleofuncional são: item (i), sistema (S), ambiente (A), propósito (P), comportamento (C) e teoria (T). Como vimos, Wimsatt discute amplamente essas variáveis, mas, de um modo resumido, podemos dizer o seguinte: i é o item em análise ou o objeto de explicação, S é o sistema ao qual i pertence, A é o ambiente no qual S está inserido, P é o propósito (ou fim) pelo qual i existe, B é o comportamento diretamente associado com i que estamos considerando, e T são teorias científicas gerais que estabelecem os limites das explicações teleofuncionais. Uma explicação teleofuncional seria, nesse contexto, a soma dessas variáveis.

Cabe aqui utilizarmos novamente do exemplo do coração para tornar isso mais claro. Nesse caso, temos que i é o tipo coração humano. S seria, nesse contexto, o corpo humano

e *A* seria o ambiente evolutivo normal no qual o tipo coração humano foi selecionado. O propósito *P*, aspecto central para as explicações teleofuncionais, é o de bombear o sangue no organismo de alguns animais de tal modo que seu organismo possa funcionar corretamente. Por fim, as variáveis *C* e *T*, embora um pouco menos intuitivas, estão associadas a um contexto de desambiguação mais geral das explicações teleofuncionais. *C*, por exemplo, é utilizada em casos nos quais o item *i* em questão possua mais de uma função. Neste caso, devemos determinar um comportamento *C* em específico que nos permita diferenciar qual função é o objeto de nossa explicação. *T*, por outro lado, serve para restringir a nossa explicação a um contexto de coerência com outras teorias científicas. Tendo essas considerações em mente, podemos finalmente apresentar o cálculo relativo a uma explicação teleofuncional do coração nos seguintes termos:

> F (função) = i (coração) + S (seres humanos) + P (bombear o sangue no organismo de alguns animais de tal modo que seu organismo possa funcionar corretamente) + B (determinados movimentos do órgão) + E (em determinadas condições normais especificadas pela história evolutiva) + T (teorias mais gerais da biologia)

Antes de concluirmos essa parte, um esclarecimento é necessário. Note que quando dizemos que a função do coração é a de bombear o sangue, dizemos que a função *F* dos corações equivale a própria variável *P*. Isso parece ser problemático, visto que parece haver uma sobreposição de *P* em relação às outras variáveis. Poderíamos, em um certo sentido, dizer que isso realmente acontece, mas seria preciso qualificarmos o sentido em que dizemos que *P* se sobrepõe às outras variáveis. Quando dizemos que a função *F* equivale a *P*, estamos expressando um caráter muito importante e ao mesmo tempo distintivo do tipo de explicações que consideramos aqui. Em outras palavras, ao dizer que *F* equivale a *P*, estamos assumindo uma relação estreita entre função e teleologia. Essa equivalência entre *F* e *P*, entretanto, é uma equivalência que se dá somente no âmbito linguístico. É certamente mais econômico dizer que "A função *F* é *P*" do que apresentar o "cálculo" que esquematizamos acima, mas isso não implica a desqualificação das outras variáveis. Essas variáveis, como espero ter deixado claro nessa seção, são extremamente importantes para que possamos apresentar as explicações teleofuncionais com o rigor necessário de uma explicação científica. Por fim, tendo discutido os aspectos principais das funções etiológicas e das explicações teleofuncionais, vejamos agora como elas

se relacionam com o funcionalismo em filosofia da mente.

3.3 Funções etiológicas e teleofuncionalismo

Uma posição bastante conhecida na filosofia da mente é o eliminativismo. De modo geral, os eliminativistas defendem a eliminação do vocabulário de senso comum no que se refere às entidades mentais. De acordo com o eliminativismo, as entidades que o vocabulário mental de senso comum alude são muito imprecisas para descrever o que acontece em nossos cérebros. Daniel Dennett (1987) propõe o termo *folk psychology* (psicologia de senso comum) para descrever a posição teórica que aceita a caracterização de senso comum dos estados mentais. Entre os termos mais usados, podemos destacar termos como crenças, desejos, amor, medo, etc. como exemplos paradigmáticos dessas entidades *folk*.

Os eliminativistas são simpáticos aos avanços da neurociência e acreditam que uma ontologia definitiva da mente deve ser dada nos termos dessa ciência. Assim, um eliminativista não acredita que o vocabulário da psicologica de senso comum terá sucesso como ponto de partida na investigação científica da mente. Em seu livro *Content and Consciousness*, de 1969, Dennett apresenta uma distinção bastante útil para se utilizar no estudo da mente. De acordo com Dennett, podemos conceber os estados mentais a partir de duas perspectivas distintas: (i) a perspectiva que preza pelo sujeito; e (ii) a perspectiva que preza pelas partes constituintes desse sujeito. De modo mais preciso, Dennett diz que no primeiro caso adotamos uma perspectiva que preza pela análise no *nível pessoal*, enquanto que no segundo adotamos uma análise que parte do *nível sub-pessoal*.

Essa distinção é particularmente interessante para a neurociência e a inteligência artificial. Os termos de senso comum como crenças e desejos dependem da realização de processos bastante complexos que envolvem o funcionamento coordenado de várias partes distintas do cérebro de uma pessoa. Assim, um cientista que investiga as bases neurofisiológicas de uma crença pode, primeiro, postular a existência da crença no nível pessoal (como pertencente ao sujeito) e então analisar como ela é possível a partir do funcionamento das estruturas cerebrais (uma análise em nível sub-pessoal). Assumindo que uma visão funcionalista sobre a natureza da mente esteja correta, poderíamos então aplicar um processo de "engenharia reversa" no es-

tudo do cérebro, o que permitiria a identificação de estruturas mais simples que poderiam ser reproduzidas em sistemas artificiais até que tenhamos como resultado um sistema que, no nível pessoal, pudesse ter uma crença ou um desejo.

Um estado mental é analisado, portanto, em termos dos processos subjacentes à sua realização. Essa transição no nível de análise permite ao teórico compreender de modo mais preciso o processo em questão, de modo que termos como crenças ou desejos possam ser explicados em termos funcionais mais completos. O problema que surge ao aceitarmos essa distinção de Dennett é que ela torna o eliminativismo uma postura atrativa na filosofia da mente. Isso se torna claro na medida em que parece arbitrário optar pela ontologia descrita no nível de pessoal como ontologia definitiva da mente. Não poderíamos pensar uma ontologia dos estados mentais no nível sub-pessoal? Não seria isso, inclusive, algo desejável para um estudo científico da mente?

Uma questão que se coloca é como essa discussão ontológica se relaciona ao problema dos $qualia$ que estamos discutindo. A resposta é que os $qualia$, assim como crenças e desejos, também fazem parte do vocabulário de senso comum da mente. Em outras palavras, a ontologia dos $qualia$ descreve o modo em que os concebemos os aspectos qualitativos de nossos estados mentais a partir do nível pessoal. Desse modo, se os $qualia$ fazem parte do vocabulário de senso comum, então eles também podem ser objeto do eliminativismo.

Nessa parte, formularei o teleofuncionalismo de modo mais preciso e mostrarei que ele se compromete com uma versão epistêmica do eliminativismo. Com isso, pretendo dizer que ao assumirmos que os $qualia$ podem ser estudados a partir da perspectiva de nível sub-pessoal, podemos justificar nossas teses em relação a eles no nível pessoal. Não pretendo tomar parte no debate ontológico sobre qual nível descreve melhor a natureza dos estados mentais, mas sim mostrar que a análise no nível sub-pessoal é capaz de captar peculiaridades epistêmicas que não são vistas em uma análise de nível pessoal.

3.3.1 A concepção tradicional de *qualia*

Antes de prosseguir, retomemos a caracterização dos $qualia$ dada no início dessa dissertação. Como vimos anteriormente, muitos dos estados mentais aos quais estamos sujeitos ao longo

de nossa vida possuem aspectos qualitativos. Quando dizemos que um estado mental possui um aspecto qualitativo, queremos dizer que ele possui um aspecto que o torna distinto de outros estados mentais no que se refere a como é estar sujeito a esse estado mental específico. Casos como a dor, a experiência visual do vermelho, ou a doçura de um alimento são casos paradigmáticos desses estados. Como observamos intuitivamente, cada uma dessas experiências tem diferentes aspectos, isto é, o aspecto *dolorido* (*painfulness*) da dor, a *doçura* do mel e a *vermelhidão* de uma maçã.

Em uma definição mais precisa, podemos descrever os $qualia$ da seguinte forma:

(i) são propriedades intrínsecas;
(ii) são propriedades subjetivas e inefáveis; e
(ii) são propriedades brutas ou monádicas.

Compreendamos cada um desses pontos com mais precisão. Quando dizemos que os $qualia$ são intrínsecos, fazemos referência ao modo especial pelo qual conhecemos um $quale$ particular. O $quale$ da dor, por exemplo, é uma propriedade intrínseca à minha experiência porque não preciso realizar nenhum processo de inferência para concluir que sinto uma dor. Basta que seja o caso que eu tenha uma experiência de dor que essa experiência será intrinsecamente uma experiência com o $quale$ da dor.

No caso de (ii), $qualia$ são ditos inefáveis e essencialmente subjetivos porque sua existência não pode ser observada a partir do ponto de vista de terceira pessoa. Quando sinto uma dor, somente *eu* posso sentir essa dor. Isso indica que o $quale$ da minha experiência de dor é subjetivo porque nenhuma outra pessoa tem acesso a ele. Essa caraterização é bastante expressiva no caso da neurociência: parece pouco provável que as interações eletroquímicas que ocorrem no cérebro quando experimento um pouco de mel possam revelar aos cientistas *como é* ter a experiência de experimentar o mel.

Por fim, em (iii) dizemos que os $qualia$ são propriedades brutas ou monádicas porque elas não parecem ser passíveis de redução a qualquer outro elemento dos estados mentais. Em outras palavras, os $qualia$ são as unidades ontológicas mais básicas quando se tratam dos aspectos qualitativos dos estados mentais.

O modo de conceber os $qualia$ a partir de (i)-(iii) é característico da concepção de senso comum da mente. Embora seja uma definição com certo grau de precisão, ela parece enfrentar

dificuldades quando é empregada como objeto de estudo das ciências naturais. Isso ocorre porque esse esquema conceitual não parece se adequar àquilo que o estudo dos estados mentais a partir do cérebro nos revela. Aqui novamente nos encontramos com um difícil dilema: ou salvamos a concepção de senso comum dos $qualia$ ou assumimos a insuperável limitação da ciência para estudar a mente.

Na próxima parte, tentarei mostrar que a concepção tradicional dos $qualia$ (a concepção de senso comum) é limitada e que deve dar lugar a uma concepção mais sofisticada desenvolvida a partir do nível sub-pessoal. Com isso, não pretendo argumentar em favor de uma tese reducionista, mas sim em favor de uma revisão conceitual da noção de $qualia$ e o modo em que entendemos essa noção. Ao final, essa revisão conceitual será bastante útil para formularmos a teoria epistêmica dos $qualia$.

3.3.2 É a concepção tradicional dos *qualia* confiável?

A oposição entre "objetivo" e "subjetivo" aponta um problema ontológico no que se refere a uma possível explicação científica dos $qualia$. Em outras palavras, como explicar a natureza de propriedades subjetivas em termos objetivos? Uma proposta que assuma tal tarefa parece estar fadada ao fracasso. Neste ponto, duas possibilidades surgem para abordar o problema: ou assumimos que os $qualia$ não são objetos de explicação científica, ou então procuramos reformular nosso quadro conceitual para poder acomodá-los.

Antes de prosseguir, gostaria de tecer alguns comentários referentes à relação entre a reformulação do quadro conceitual que usamos para estudar os $qualia$ e a questão da redução desses últimos a entidades físicas. Acredito que embora ambas as questões estejam relacionadas, elas são independentes. Quando falamos de uma reformulação no modo em que compreendemos a noção de $qualia$, há claramente uma preocupação em tornar o problema mais tratável para a ciência. O que deve estar claro, no entanto, é que essa preocupação não implica assumir uma tese reducionista sobre a natureza dos $qualia$. Isso quer dizer que o fato de uma entidade ser passível de explicação científica não implica que ela seja redutível àquilo que uma ou mais ciências estabelecem como unidades mais básicas do mundo. Nesse sentido, a reformulação da concepção tradicional de $qualia$ que proponho aqui parte do pressuposto que há uma correlação

entre estados mentais e estados cerebrais. A natureza dessa correlação, no entanto, não é importante para os nossos propósitos aqui. Assim, a discussão que apresento a seguir não pretende sugerir uma tese ontológica, mas somente uma tese epistêmica que diz respeito aos $qualia$ e como eles se relacionam com processos cerebrais.

Comecemos a discussão analisando criticamente a concepção tradicional de $qualia$. Como vimos na seção anterior, os $qualia$ são definidos por três características principais: (i) eles são intrínsecos; (ii) são subjetivos e inefáveis; e (iii) são brutos ou monádicos. Essa concepção, como mencionei rapidamente, confronta a caracterização de mente apresentada pela neurociência. Em seu artigo *Eliminative materialism and propositional attitudes*, Paul Churchland (1981) sugere que os termos da psicologia de senso comum sejam substituídos por termos apresentados por teorias mais sofisticadas da neurociência. Para Churchland, a psicologia de senso comum é um modo muito grosseiro (e em alguns casos equivocado) de descrever aquilo que acontece em nossos cérebros.

O argumento de Churchland em favor da eliminação da psicologia de senso comum é um argumento histórico. Churchland diz que a psicologia de senso comum consiste numa teoria primitiva que será eventualmente eliminada por uma ciência futura, assim como ocorreu com as teorias do flogisto e do calórico. Nesses casos, entidades que os cientistas acreditavam existir e ter caráter explicativo (flogisto e calórico) foram simplesmente eliminados da ontologia da ciência depois que novas teorias mais sofisticadas foram apresentadas para explicar os fenômenos que essas entidades lidavam.

Outra proposta de eliminação da psicologia de senso comum, e essa centrada particularmente no caso dos $qualia$, é discutida por Dennett (1988) em *Quining qualia*. Nesse artigo, Dennett descreve vários casos hipotéticos que geram dúvidas sobre a plausibilidade do vocabulário tradicional em descrever os $qualia$. Tendo em vista o enfoque direto dado por Dennett, consideraremos essa discussão com mais detalhes. Dennett argumenta que falar dos $qualia$ como propriedades intrínsecas e monádicas não faz sentido quando consideramos essa caracterização em experiências específicas. Considere o caso dos enólogos. Na tentativa de explicar por que esses indivíduos têm muito mais sucesso em identificar propriedades do vinho do que nós, pessoas não-treinadas, dizemos que a experiência qualitativa que esses indivíduos

possuem é significativamente distinta da nossa. Com isso, dizemos que os enólogos são capazes de identificar a composição química do vinho, sua textura, a época em que foi produzido, etc. porque seus $qualia$ são distintos dos nossos $qualia$. O problema com esse caso é que se essa explicação estiver correta, então os $qualia$ não podem ser intrínsecos. Isso se dá porque algumas das propriedades dos $qualia$ que experimentamos ao tomar o vinho simplesmente não são reveladas a nós. Para serem identificadas, elas dependem de alguma forma de conhecimento proposicional e algum processo de inferência (tal como o faz o enólogo a partir de seu conhecimento sobre vinho). O mesmo pode ser dito do aspecto (iii). Se podemos distinguir aspectos mais básicos e elementares da experiência de tomar o vinho, então isso quer dizer que a experiência qualitativa não é homogênea, mas sim composta de partes distintas. A experiência qualitativa do enólogo deixa de ser um bloco único, sendo composta agora de várias partes distintas, o que o permite identificá-las.

Até aqui apresentei os argumentos de Dennett contra (i) e (iii). Mas o que dizer de (ii)? Parece que as considerações feitas acima não afetam o aspecto (ii) da definição tradicional de $qualia$ como propriedades subjetivas e inefáveis, o que, de certo modo, é o que fundamenta o problema epistêmico. Considere o caso de uma dor de cabeça. Quando digo que tenha uma dor de cabeça, essa dor é somente minha no sentido em que ninguém mais pode senti-la. O termo que uso para me referir ao $quale$ dessa dor de cabeça é, nesse sentido, um termo com um significado subjetivo, uma vez que apenas eu posso sentir e falar sobre os aspectos qualitativos da minha dor. Mais ainda, parece claro que quando tenho uma dor de cabeça, ela tem sua base em algum processo cerebral, caso contrário os analgésicos não teriam efeito. Como a dor tem essa dimensão física, sentimo-nos inclinados a crer que as dores que sentimos são da mesma natureza qualitativa.

Agora considere o caso de um viajante do tempo vindo do ano 2375. Considere ainda que a neurociência nesse ano seja tão avançada que se torna comum para esse indivíduo dizer que ele possui uma estimulação de fibras-C no cérebro quando ele tem uma dor de cabeça. Considere, entretanto, que esse viajante desconheça a história da neurociência, de modo que ele estranha a nossa insistência em aceitar que aquilo que nós chamamos de dor de cabeça

é somente a estimulação de fibras-C no cérebro[4]. Se o viajante do tempo pudesse ler essa dissertação, mais especificamente a parte em que definimos as experiências conscientes como subjetivas, não seria espantoso se ele protestasse. Para ele, é perfeitamente claro que o que ele se refere como "dor" não é algo subjetivo: a sua dor é simplesmente o resultado da estimulação de fibras-C em seu cérebro, sendo elas perfeitamente observáveis de um ponto de vista de terceira pessoa. Como aponta Paul Churchland (1989, p. 151), o que caracteriza uma explicação de primeira pessoa é apenas a familiaridade da terminologia usada e a espontaneidade do uso de um esquema conceitual específico. Nesse sentido, não há nada que nos impeça de descrever nossos estados mentais subjetivos em termos objetivos:

> A neurociência pode parecer incapaz de nos dar uma explicação puramente em "terceira pessoa" da mente, mas apenas a familiaridade e a espontaneidade de respostas conceituais são requeridas para tornar essa explicação de "primeira pessoa". O que torna uma explicação de primeira pessoa não é o conteúdo dessa explicação, mas sim o fato de que podemos usá-la como veículo de conceitualização espontânea na introspecção e na autodescrição. (CHURCHLAND, 1989, p. 151)

Parece concebível, portanto, que no caso do viajante do tempo, a descrição de seus $qualia$ não seja mais subjetiva, mas sim objetiva. Isso seria um indicativo do fato de que o aspecto subjetivo dos $qualia$ não é uma propriedade essencial desses últimos, mas somente do modo em que aprendemos a falar sobre eles.

Para concluir essa parte, um último comentário é necessário. Os argumentos apresentados aqui, tal como destaquei inicialmente, são argumentos que visam minar uma determinada concepção de $qualia$. Nas próximas seções, argumentarei que o tipo de identidade pressuposta pelo teleofuncionalismo não se compromete com qualquer domínio ontológico, seja ele mental ou físico. Nesse sentido, a conclusão a ser tirada dessa discussão é somente que a concepção tradicional é bastante problemática e que ela deve ser superada para avançarmos em relação ao problema epistêmico.

[4]Esse exemplo se encontra formulado de outro modo em Rorty (1973).

3.3.3 Teleofuncionalismo e *qualia*

Até aqui, discuti essencialmente o problema dos *qualia* e os aspectos fundamentais das explicações funcionais. Agora utilizarei dessas ferramentas conceituais para reformular o funcionalismo e relacioná-lo, na perspectiva dessa reformulação, com o problema dos $qualia$. A discussão feita nos capítulos anteriores mostrou que o funcionalismo consiste em uma doutrina geral sobre a natureza da mente. Nesse sentido, diversas versões do funcionalismo podem ser formuladas, destacando diferentes comprometimentos teóricos. Como vimos, o problema epistêmico se coloca frente a essa formulação mais geral e tradicional, principalmente associada aos trabalhos de Putnam (1973). Pretendo argumentar que as dificuldades que vimos anteriormente não se colocam a uma versão mais específica do funcionalismo que se alia à noção de teleologia. Esse *funcionalismo teleológico*, ou simplesmente *teleofuncionalismo*, servirá de base para considerarmos o problema epistêmico daqui em diante.

Para entender melhor no que consiste o teleofuncionalismo, considere o tipo de análise funcional proposta por Cummins (1975). Cummins argumenta que explicar a função de um órgão biológico ou de um artefato depende de uma análise de cima pra baixo (*top-down*), ou, para usar outro termo, uma *análise de decomposição*. Nesses casos, atribuímos uma função Y para um sistema S e então decompomos a tarefa de S de realizar Y em tarefas mais simples do tipo $y_1 \ldots y_n$ e atribuímos elas aos subsistemas de $s_1 \ldots s_n$ de S. Desse modo, a tarefa de S de realizar Y pode ser analisada de modo mais simples a partir do estudo de suas partes.

O teleofuncionalismo, tal como o entendo, faz uso do tipo de análise funcional de Cummins, embora não partilhe da noção de função de Cummins. Essa aparente tensão será esclarecida em seguida. Na análise funcional, portanto, estados mentais são considerados entidades complexas que são melhor compreendidos pela análise de seus mecanismos subjacentes. Assim, quando investigamos a natureza de um estado mental, deixamos de vê-lo como um bloco homogêneo (nível pessoal) para analisá-lo a partir de suas partes componentes (nível sub-pessoal). Ainda nessa linha de raciocínio, William Lycan (1995) utiliza da metáfora de "instituições"no cérebro para ilustrar a ideia por trás do teleofuncionalismo. De acordo com essa imagem, as "partes componentes" de um estado mental seriam associadas a partes específicas do cérebro (as instituições), de modo que o funcionamento conjunto delas torna possível a existência de

um todo que é o estado mental considerado a partir do nível pessoal.

Dois pontos nos interessam na analogia de Lycan (1995): primeiro, a capacidade de dividir o trabalho explicativo (*explanatory burden*) referente à natureza de um estado mental sem perder de vista a perspectiva de primeira pessoal. Em termos epistêmicos, o teleofuncionalismo aceita a complexidade de um estado mental no nível pessoal, mas busca explicar essa complexidade em termos sub-pessoais. O segundo ponto se refere ao modo em que a divisão de instituições dentro do cérebro é feita. Aqui podemos entender claramente porque o teleofuncionalismo utiliza da análise funcional de Cummins, mas não partilha de sua noção de função. Como sugere Ruth Millikan (2002), uma análise funcional nos moldes propostos por Cummins só é possível quando estabelecemos um contexto etiológico por trás dela. Para Millikan, só podemos dizer que a função de um traço biológico t é f se soubermos, de um modo geral, qual a função g do sistema que t faz parte. Assim, quando tratamos tanto de casos biológicos quanto de artefatos, essa contextualização só pode ser dada em termos de história de seleção[5].

O que podemos extrair do segundo ponto é que se a análise funcional depende da história seletiva, então ela depende intimamente do estabelecimento de um fim (seguindo os pressupostos etiológicos) para um estado mental em nível pessoal. Esse é o caso porque somente assim funções mais básicas podem ser atribuídas aos componentes do nível sub-pessoal. Encontramos, portanto, duas propriedades essenciais para uma teoria funcionalista na formulação do teleofuncionalismo. Primeiro, ele permite dividir o esforço epistêmico dentro dos preceitos do funcionalismo, e segundo, o que permite essa divisão é o uso de noções etiológicas, o que ressalta a importância da concepção etiológica.

A sugestão óbvia por trás dessa formulação é que, assim como qualquer outro estado mental passível de uma análise teleofuncional, a estratégia de decomposição também pode ser usada no caso dos $qualia$. Como vimos, os $qualia$ são partes bastante complexas do nossos estados mentais que embora sejam característicos de uma análise no nível pessoal, podem se beneficiar de uma explicação em termos sub-pessoais. A motivação por trás dessa sugestão é bastante simples: uma vez que reconhecemos o problema dos $qualia$ como consequência da concepção tradicional, um estudo que não se paute nessa concepção pode se mostrar frutífero

[5] Argumento mais detalhadamente em favor dessa tese em Sant'Anna (2014).

na resolução ou na compreensão dessas questões.

3.3.4 Eliminativismo e teleofuncionalismo

Mencionei no início dessa seção que a distinção entre nível pessoal e nível sub-pessoal consiste em um passo na direção de uma postura eliminativista. Na verdade, o teleofuncionalismo aceita essa distinção e também assume uma versão da tese eliminativista. Esse eliminativismo, no entanto, precisa ser caracterizado para evitar possíveis desentendimentos. Nesta parte, tentarei mostrar por que o teleofuncionalismo consiste em uma ameaça para a concepção tradicional de $qualia$ e que tipo de tese eliminativista ele defende.

Como vimos, Dennett (1988) argumenta que a concepção tradicional de $qualia$ não é adequada. Para ele, essa concepção não é coerente com aquilo que conhecemos do cérebro e até mesmo com o modo que descrevemos nossas experiências. Nessa perspectiva, Dennett sugere a eliminação do modo tradicional de caracterizar os $qualia$. O primeiro argumento apresentado por Dennett se concentra na impossibilidade da concepção tradicional de $qualia$ em explicar os aspectos qualitativos em toda sua complexidade. Isso se torna claro no caso hipotético de Chase e Sanborn apresentado por Dennett. Chase e Sanborn são dois avaliadores de café profissionais. Quando ambos tomam o mesmo café, eles fazem relatos baseados em sua introspecção. O problema é que é possível, como ocorre em muitos casos do cotidiano, que Chase e Sanborn façam relatos opostos de suas experiências. Enquanto um pode fazer uma avaliação positiva, o outro pode fazer uma avaliação extremamente negativa. Dennett argumenta que essa diferença não pode ser explicada pela concepção tradicional, uma vez que ela assume que os $qualia$ são intrínsecos, subjetivos e inefáveis. A saída mais plausível seria dizer que os $qualia$ de Chase e Sanborn são diferentes, mas isso implicaria o problema epistêmico. O desafio é, portanto, encontrar uma saída para essa dificuldade.

A solução que sugiro é utilizar da distinção entre nível pessoal e nível sub-pessoal de análise. Como vimos, a concepção tradicional é dada a partir do nível pessoal. Assim, um $quale$ particular é visto como uma unidade básica que não pode ser dividida para ter sua complexidade explicada. Isso ocorre porque essa análise é dada essencialmente pela introspecção. Se Dennett estiver correto, no entanto, a introspecção não fornece um bom caminho de investigação para

os $qualia$.

Com essas dificuldades em mente, uma alternativa é apelar para uma análise fundamentada nas pressuposições do nível sub-pessoal de análise. Uma explicação no nível sub-pessoal não toma como ponto de partida a perspectiva do agente para analisar os estados mentais, mas, ao contrário, propõe a divisão das entidades consideradas brutas ou monádicas no nível pessoal em subsistemas no nível sub-pessoal. Em outras palavras, as peculiaridades de um fenômeno como a dor são associadas a partes específicas do cérebro — isto é, uma associação no nível sub-pessoal — de modo que diferentes tipos de atribuições funcionais são feitas em cada subdivisão. Essa divisão epistêmica, por sua vez, torna o fenômeno mais maleável, uma vez que possibilita a decomposição de um fenômeno complexo em fenômenos mais simples. Assim, quanto mais dividimos essas entidades inicialmente monádicas, atribuindo a elas funções específicas mais simples, menor será a complexidade exigida para que essa função seja realizada. Nesse sentido, a subsequente divisão de uma única função em sub-funções, e assim sucessivamente, levará a um cenário ideal no qual as interações funcionais poderão ser descritas em termos de 0's e 1's. Desse modo, a quebra de um fenômeno complexo e inicialmente considerado bruto, homogêneo ou monádico acomoda um grau de complexidade de análise muito mais detalhado do que uma análise restrita ao nível pessoal, o que aumenta significativamente nossas capacidades explicativas em relação ao fenômeno estudado. Assim, é seguro concluir que se pretendemos apresentar uma análise que faça justiça à complexidade dos fenômenos qualitativos, então devemos abandonar a concepção tradicional.

Parece claro, portanto, que as críticas de Dennett apontam para uma análise que adota os pressupostos teleofuncionalistas. De acordo com Dennett, essa concepção nos levaria a uma tese eliminativista sobre os $qualia$. Nessa versão, o eliminativismo assume que a concepção tradicional não descreve bem o que ocorre em nossos cérebros. Essa imprecisão, por sua vez, desqualifica os $qualia$ — tradicionalmente concebidos — como um conceito seguro para tratar dos aspectos qualitativos de nossas experiências conscientes.

Temos claro agora o sentido em que o teleofuncionalismo propõe uma tese eliminativista. Em outras palavras, o teleofuncionalismo advoga em favor da eliminação metafísica da noção dos $qualia$, e não propriamente uma eliminação ontológica. Nesse sentido, o eliminati-

vismo que defendo não nega que temos dor quando tocamos uma chapa quente, mas sim que o modo que descrevemos essa experiência, e mais ainda, o modo que teorizamos sobre ela, está equivocado. Por esse motivo, o eliminativismo metafísico não se compromete com teses ontológicas sobre os $qualia$. Ele aceita a existência desses últimos, mas se mantém neutro quanto à sua relação com o mundo físico. A caracterização dos $qualia$ dada pelo teleofuncionalismo é puramente funcional, o que é refletido no argumento em favor da análise em nível sub-pessoal.

Mas isso deixa uma questão em aberto: é possível falar de identidade no sentido ontológico dentro do teleofuncionalismo? Esse tipo de proposta reducionista coloca um problema para o teleofuncionalismo porque estados mentais parecem ter uma relação bastante íntima com estados cerebrais. Assim, se o projeto reducionista tiver sucesso, poder-se-ia ganhar em economia ontológica. Similarmente, eliminativistas clássicos como Churchland e Churchland (1996) simplesmente negam a existência dos fenômenos mentais, o que também resulta numa economia ontológica.

Acredito que o teleofuncionalismo e o eliminativismo, como formulados aqui, não precisam se comprometer com nenhum desses extremos. Em outras palavras, não é o caso que ou reduzimos os estados mentais ou simplesmente os eliminamos. Essa disjunção é falsa, uma vez que o teleofuncionalismo não se compromete com nenhuma dessas teses. Com isso, pretendo enfatizar uma diferente noção de identidade expressa pela *teoria da identidade heurística* (TIH) defendida por McCauley e Bechtel (2001). De acordo com Bechtel e McCauley (2001), a TIH sustenta que uma identidade deve ser vista como uma premissa e não uma conclusão de uma investigação científica. Isso quer dizer que a identidade dos estados mentais e estados cerebrais tem, em seu princípio, um valor heurístico para a investigação. Assim, em oposição ao reducionista tradicional, a identificação de um estado mental m com um estado cerebral c não é uma consequência de uma teoria neurocientífica bem desenvolvida, mas sim o que torna possível realizar a investigação a partir desses termos.

Compreendamos esse ponto com mais detalhes. McCauley e Bechtel argumentam que identidades no nível cerebral não precisam ser identidades detalhadas e bem delimitadas (*fine-grained identities*). O que importa numa identidade no nível cerebral é que ela estabeleça uma agenda de pesquisa no sentido em que, sendo uma área do cérebro identificada com uma função,

experimentos podem ser realizados para testar se essa identidade é teoricamente razoável ou não. Os experimentos, no entanto, não devem ser realizados apenas no domínio de uma única disciplina científica. Em outras palavras, um pesquisador que aceita a TIH não deve testar sua hipótese de identidade apenas nos domínios da neurociência ou da psicologia. O tipo de investigação proposto por McCauley e Bechtel é uma que integra diversos domínios de investigação. Isso se justifica na medida em que diferentes disciplinas têm interesses em um mesmo fenômeno, mas cada uma tem métodos e pressupostos teóricos diferentes. Essa discrepância gera experimentos diferentes em cada domínio e que podem elucidar dificuldades existentes em outras áreas. Para McCauley e Bechtel, adotar a TIH gera uma espécie de pluralismo explanatório. No caso dos estados mentais, estes são investigados por múltiplas áreas de investigação, o que gera uma visão mais completa e frutífera do fenômeno que é objeto de estudo.

Algo importante a se notar é que o tipo de teoria que resulta do uso da TIH é o que Darden e Maull (1977) chamam de *teorias interfield*. De acordo com Darden e Maull, teorias *interfield* surgem quando "duas áreas compartilham um interesse em explicar diferentes aspectos de um mesmo fenômeno" (1977, p. 133). Nesse sentido, diferentes aspectos dos estados mentais podem ser investigados por diferentes domínios de investigação. Isso é particularmente importante para o que podemos chamar de ciência cognitiva, na qual diversos aspectos da mente são estudados por diferentes áreas, como a neurociência, a psicologia, a filosofia, a antropologia, a biologia, a inteligência artificial, etc.

Como Darden e Maull enfatizam, essa abordagem permite ao investigador responder questões que parecem intratáveis a partir do uso de conceitos e técnicas de outros domínios. Elas citam, como exemplo, a contribuição mutual que a citologia e a genética forneceram para a criação de uma teoria dos cromossomos. Para resumir, podemos dizer que uma teoria *interfield* é gerada quando:

> O conhecimento de fundo indica que existem relações entre áreas, quando as áreas compartilham de um interesse em explicar um mesmo fenômeno, e quando questõs surgem sobre um fenômeno em determinada área que não podem ser respondidas com as técnicas e conceitos daquela área. (DARDEN; MAULL, 1977, p. 134)

Um questionamento interessante a ser feito é como a relação entre as áreas participantes em uma teoria *interfield* pode ser explicadas. Em outras palavras, como podemos elencar os conceitos e assunções teóricas de uma área com outras áreas? Darden e Maull (1977) fornecem uma análise de como isso é possível no caso da citologia e da genética, mas para nossos propósitos, deixaremos esse caso de lado. Proponho, em contrapartida, considerarmos diretamente o caso da integração entre psicologia e neurociência, um caso que é de nosso interesse primário.

Técnicas de radiografia como o fMRI (*Functional Magnetic Resonance Imaging*) e o PET (*Positron Emission Tomography*) são muito comuns nos dias de hoje. Um de seus usos consiste em ajudar a estabelecer a função de determinadas regiões do cérebro. Um exemplo simples é capaz de ilustrar isso. Quando a glicose e o oxigênio são metabolizados no cérebro, essa metabolização se dá de modo proporcional ao uso de determinada região do cérebro. No caso do PET, insere-se oxigênio radioativo na corrente sanguínea de um paciente e pede-se para que ele realize algum tarefa mental. Enquanto o sujeito realiza a tarefa, o exame radiográfico é feito. É possível observar no resultado a presença do oxigênio radioativo nas áreas que foram utilizadas quando o indivíduo estava realizando aquela tarefa. Essa é uma primeira indicação de que as áreas indicadas no exame são funcionalmente responsáveis pela habilidade do indivíduo em realizar aquela tarefa.

É importante notar que quando o indivíduo realiza uma operação mental, ele recebe instruções de como proceder no nível psicológico. Em outras palavras, o sujeito é instruído nos termos da psicologia de senso comum apresentados acima. Nesses experimentos, o indivíduo usualmente precisa *descrever* algo, *identificar* algo, ou até mesmo *decidir* fazer algo. Todos esses termos compõem o vocabulário da psicologia de senso comum. Isso mostra que a psicologia de senso comum é importante pelo menos no início de uma investigação mais minuciosa sobre as funcionalidades do cérebro.

Isso nos permite ver que a agenda de pesquisa da psicologia e da neurociência parecem estar relacionadas de algum modo. Mas a questão de Darden e Maull (1977) se colocar novamente: como essa relação é possível? Aqui podemos identificar a importância da biologia evolutiva e da concepção etiológica de funções para o teleofuncionalismo. Essa relação é

explorada por Jennifer Mundale e William Bechtel (1999). Eles sugerem que a integração entre psicologia, neurociência e biologia evolutiva se dê a partir da noção etiológica de função. Consideremos esse ponto com mais detalhe.

A assunção geral que sustenta esse projeto é que o cérebro é, assim como grande parte dos órgãos biológicos, o resultado de um processo de seleção adaptativo, isto é, a seleção natural. Naturalmente, o cérebro e suas partes componentes têm funções que foram executadas no passado e que beneficiaram os indivíduos que podiam executá-las. Essas funções são os motivos pelos quais essas áreas foram selecionadas, explicando o porquê de elas existirem. Os termos da psicologia de senso comum que usamos para descrever nossos estados mentais são, portanto, funções que foram selecionadas para serem passadas pelas gerações, uma vez que elas consistem em grande vantagem para a nossa sobrevivência. A capacidade de discriminar um inimigo de um amigo é essencial, o que exige a seleção de mecanismos que possibilitem a realização dessas funções.

Torna-se claro, portanto, que uma vez que consideramos nossas habilidades psicológicas de um ponto de vista evolutivo, podemos então justificar (em termos evolutivos) a atribuição de funções para as áreas ativadas no PET. Nesse sentido, aquilo que aprendemos através da biologia evolutiva deve modelar nossa compreensão das nossas habilidades psicológicas no mesmo sentido em que experimentos da psicologia desvendam fatos importantes sobre nossa história evolutiva. Além disso, nada exige que nos restrinjamos a essas áreas. Na verdade, a neurociência também investiga esses fenômenos de uma perspectiva diferente. Essas investigações com técnicas diferentes e metodologias diferentes também tem importantes implicações para as duas áreas mencionadas acima[6].

A discussão das teorias *interfield* foi necessária para apresentar com mais precisão a posição ontológica minimalista do teleofuncionalismo. É possível justificar agora que o teleofuncionalismo não se compromete com nenhum domínio ontológico das ciências empíricas. Isso ocorre porque, como espero ter mostrado, a identidade da psicologia de senso comum com outra teoria é uma identidade que tem valor heurístico. Assim, não precisamos escolher um

[6]Mundale e Bechtel (1999) apresentam uma análise de como a integração de áreas é possível no caso da nossa capacidade cognitiva de aprendizado. Não tratarei dos detalhes desse caso aqui, uma vez que o objetivo da discussão era somente mostrar como teorias *interfield* podem ser geradas no estudo da mente e como a noção de função etiológica é importante para sua composição.

domínio correto para estabelecer a ontologia do mental, uma vez que cada área de investigação pode contribuir para a pesquisa utilizando-se de suas ontologia e metodologia próprias. Isso não significa negar a possibilidade do surgimento de uma nova área resultante dos esforços comuns feitos pelas áreas contribuintes de uma teoria *interfield*. Tal cenário poderia gerar uma nova ontologia com novas técnicas para essa nova área. Isso pode ser, de fato, o futuro da ciência cognitiva, mas essa ainda é uma questão em aberto. Desse modo, o que quero enfatizar é que na ausência de uma nova área de investigação, o teleofuncionalismo não se compromete exclusivamente com nenhuma ontologia das áreas componentes de uma teoria *interfield*.

Para concluir essa parte, um comentário final sobre o problema que a iniciou, isto é, o tipo de eliminativismo proposto aqui. Deve estar claro agora que esse eliminativismo não se confunde com as formas mais tradicionais da doutrina, isto é, o teleofuncionalismo não prega a eliminação de uma ontologia em favor de outra. A principal tese do eliminativismo é enfatizar que há uma importante distinção a ser feita entre o nível pessoal e o nível sub-pessoal de análise e que essa distinção é benéfica ao estudo da mente. Isso não quer dizer que o conceito de mente dado no nível pessoal seja descartado como propõe o eliminativista tradicional, mas que as inconsistências e as virtudes que essa concepção possui sejam integradas de maneira elucidativa com outras áreas de estudo.

3.3.5 Identidades heurísticas e $qualia$

Um último ponto ainda está aberto no que diz respeito à questão da identidade. Um problema que surge em relação à discussão sobre $qualia$ e identidade tem a ver com o fato de que os $qualia$ são usualmente identificados com os aspectos qualitativos e não com seus aspectos funcionais. Nesse sentido, a discussão acima teria sido em vão, uma vez que estávamos falando das propriedades funcionais dos $qualia$ e não da sua natureza qualitativa ou subjetiva. Tentarei mostrar que esse é um modo equivocado de pensar a questão e que conceber $qualia$ como aspectos funcionais aumenta nossa compreensão sobre sua natureza.

Um bom ponto para iniciar essa reflexão é olhar para o modo em que estabelecemos a identidade de um estado mental com aspectos qualitativos. Um modo comum de proceder é dizer que um estado mental m é uma dor sse ele tiver um aspecto *dolorido*. Assim, podemos

dizer que todas experiências doloridas são ocorrências do tipo "dor". Podemos imaginar agora uma situação na qual alguém tenha uma experiência subjetiva dolorida, mas que simplesmente não se incomode com essa experiência. Esse é o caso de pacientes que usam morfina. Nessa situação, haveria uma pessoa que experiencia os aspectos qualitativos da dor, mas que simplesmente não experiencia seu aspecto dolorido[7]. A questão que se coloca é se é razoável dizer que essa pessoa tem uma experiência de dor genuína. Em outras palavras, existem dores sem serem doloridas?

Para ver isso de modo mais claro, considere o caso de emoções ou sentimentos, como é o caso do amor, do ódio ou do medo. No caso do amor, podemos caracterizá-lo de tal modo: "X ama Y sse: (a) X sente falta de Y quando Y está longe; (b) X gosta de Y; e (c) X se sente atraído por Y". Parece razoável supor que todas as pessoas que amam — em um sentido conjugal — se enquadrariam nessa caracterização. Desse modo, não é absurdo dizer que amar requer satisfazer (a), (b), e (c). Isso, no entanto, pode gerar protesto: alguém pode dizer que o que realmente identifica a experiência de amar não é essa definição funcional, mas sim o aspecto qualitativo dessa experiência, isto é, como é estar apaixonado (*what it's like to be in love*).

Nesse contexto, a seguinte experiência de pesamento pode ser sugerida. Imagine que no futuro tenhamos uma droga conhecida como "morfina do amor" que invertesse todas as caracterizações funcionais do amor, mas que não alterasse em nada o $quale$ da experiência. Seria plausível dizer que alguém que tomou a morfina do amor ainda tem a experiência de amar? Em outras palavras, poderíamos dizer que X ama Y sem satisfazer (a), (b) e (c) acima? Ou, para tornar a questão ainda mais evidente, poderíamos dizer que X ama Y se a estrutura cognitiva de X fosse organizada do seguinte modo: (d) X não sente falta de Y; (e) X não gosta de estar com Y; e (c) X se sente inconfortável na presença de Y? Como sugere Dennett (1996), dizer que há um $quale$ que identifique essa experiência enquanto tal seria como insistir em uma explicação vitalista para os processos vitais.

Um caso que pode ser problemático é o caso das cores. Cores parecem ser imunes a caracterizações funcionais, uma vez que a inversão de espectro é perfeitamente plausível. Não tratarei dessa dificuldade aqui, visto que será objeto de investigação no próximo capítulo. Essa

[7] Ver Dennett (1981, cap. 11) para o relato de um caso real.

questão se situa no cerne do que chamo de teoria epistêmica dos $qualia$.

Para concluir, o objetivo dessa seção era questionar a nossa intuição de que nossos estados mentais com $qualia$ são identificados com o como é (*what it's like*) ter esses estados. Como espero ter mostrado, se olharmos mais atentamente para a caracterização funcional de um estado mental com $qualia$, veremos que o que torna um estado mental em um instante de tempo t_1 o mesmo tipo de estado mental em t_2 é o fato de possuírem as mesmas peculiaridades funcionais, e não o mesmo aspecto qualitativo. Essa conclusão é importante porque nos permite falar de estados mentais nos termos da biologia evolutiva. Em outras palavras, tipos de estados mentais são agora definidos por seus aspectos funcionais, o que quer dizer que podem ser objetos genuínos da seleção natural. Se isso estiver correto, então a discussão sobre o estatuto metafísico dos $qualia$ é bastante útil, uma vez que falar sobre $qualia$ exige falar de sua definição funcional.

Capítulo 4

Enfrentando o problema epistêmico

Agora temos todas as ferramentas necessárias para enfrentar o problema epistêmico. Sugeri anteriormente que olhássemos para o problema de modo similar ao problema da causalidade identificado por Hume. Mas o que isso quer dizer exatamente? Isso significa que a relação entre $qualia$ e estrutura funcional não é uma relação lógica no sentido que, dado uma estrutura funcional *F*, um sistema com $qualia$ necessariamente resultará de *F*. Argumentei que esse caso é parecido com o da causalidade porque quando dizemos que um evento E_1 é causa de um evento E_2, a relação que estabelecemos não é lógica, mas de outra natureza. Nesse sentido, assim como Kant buscou fundamentos para explicar qual a natureza dessa relação e como podemos compreendê-la, a minha proposta é fazer o mesmo com o funcionalismo e os $qualia$. Desse modo, quando olhamos para uma rosa e a vemos como vermelha, se pensarmos na formulação clássica do funcionalismo, nada nos garante que se olharmos para essa rosa amanhã, ela será necessariamente vermelha. Mais ainda, esse raciocínio pode ser estendido para um âmbito mais geral em que questionamos a presença de mente em outros seres humanos e animais. Isso parece pouco provável e pouco intuitivo, mas no papel de filósofos, devemos reconhecer que a crença sobre a continuidade e a presença dos $qualia$ — tanto em nós como em outras sistemas — se baseia num raciocínio fundado em uma base indutiva.

O mesmo ocorre no caso da causalidade. Embora as coisas queimem de fato quando as colocamos no fogo, não se segue daí que se eu colocar algo no fogo agora que esse algo irá queimar. É muito provável que isso ocorra, mas não podemos justificar devidamente nossa crença sobre esse evento se apelarmos somente para um raciocínio de natureza indutiva. Neste

ponto, para evitar possíveis equívocos, é importante notar que a analogia proposta entre ambos os casos se sustenta somente nessa dimensão, isto é, a que explicita a necessidade de fundamentar nossa crença em uma relação do mundo que não é dada em termos lógico-dedutivos. Nesse sentido, não quero sugerir que a solução para ambos os problemas sejam similares.

Tendo isso em mente, a questão que se coloca agora é: como justificar nossa crença de que certos $qualia$ são consequências necessárias de certas estruturas funcionais? Como posso dizer com certeza que meus amigos têm mente? Dado que a relação não é lógica, o problema nos deixa frente a duas possibilidades: ou apelamos para algum tipo de inferência probabilística ou então procuramos outro modo de fundamentar essa crença. Argumentei anteriormente que o primeiro caminho gera problemas caros à filosofia da mente, o que nos deixa com o segundo caminho.

Nessa seção, formularei a teoria epistêmica dos $qualia$ baseado em algumas noções da teoria evolutiva que vimos acima, mais especificamente, a noção de funções etiológicas. A teoria epistêmica se fundamenta num princípio metafísico que nos permite justificar as crenças descritas acima. Esse princípio será desenvolvido a partir da noção de *seleção*, permitindo a formulação de duas versões específicas da teoria.

4.1 Normatividade, "mau-funcionamento" e *qualia*

No capítulo 3, discuti com mais detalhes a noção de normatividade envolvida nas explicações teleofuncionais. Um conceito importante relacionado à noção de normatividade é o conceito de *mau-funcionamento* (*malfunction*). Vimos também que uma explicação teleofuncional permite estabelecer variáveis que tornam possível a realização de um "cálculo funcional". Nessa seção, utilizaremos esses elementos para lidar com o problema epistêmico dos $qualia$.

Antes de proceder, retomemos rapidamente essas noções. No capítulo 3, analisamos a proposta lógica de Wimsatt (1972) no que se refere às explicações teleofuncionais. Wimsatt apresenta cinco variáveis associadas às explicações teleofuncionais: item (i), sistema (S), ambiente (A), propósito (P), comportamento (C) e teoria (T). Brevemente, i refere-se ao objeto de explicação, S refere-se ao sistema do qual i faz parte, E refere-se ao ambiente em que S e i se encontram, P refere-se ao propósito pelo qual i existe, C refere-se ao comportamento

associado com i, e T refere-se às teorias mais gerais que delimitam o escopo de uma explicação teleofuncional.

A sugestão que faço é que consideremos i como um estado mental com $qualia$. De um ponto de vista evolutivo, podemos dizer que um estado mental é uma adaptação ao ambiente em que nossos ancestrais viveram. Essa tese geral pode ser motivada na medida em que consideramos o fato de que estados mentais servem como "janelas" para o mundo que nos permitem acessá-lo. Assim, se um estado mental serve como meio de "navegação" no ambiente que estamos inseridos, então ele pode ter vantagens evolutivas significativas. Consequentemente, estados mentais podem ser objetos de seleção natural.

Essas considerações geram alguns problemas filosóficos, uma vez que estados mentais apresentam o mundo como ele *aparece* para nós, e não como ele realmente *é*. Embora essa seja uma questão bastante importante, ela diz respeito a questões mais fundamentais sobre a natureza da percepção que não trataremos aqui. De qualquer modo, vale ressaltar que ainda que assumamos uma teoria direta ou indireta da percepção, o valor evolutivo dos estados mentais não é afetado. O que muda é somente o modo que concebemos a natureza da interação entre mente e mundo, e não propriamente a existência dessa relação. Assim, para os nossos propósitos, podemos dizer que um estado mental pode ser considerado como uma *representação* do ambiente externo. Essa é a *Tese Representacionalista* (TR) de Dretske (1995).

> Se, de acordo com a Tese Representacionalista, pensarmos fatos mentais como fatos representacionais, a qualidade da experiência, como as coisas nos aparecem no nível sensorial, é constituída pelas propriedades que respresentamos as coisas como tendo. A minha experiência de um objeto é a totalidade dos modos em que esse objeto aparece para mim, e o modo em que esse objeto aparece para mim é o modo em que meus sentidos o representam. (DRETSKE, 1995, p. 1)

Ao adotar a TR, não pretendo tomar partido no debate sobre a natureza relacional das experiências perceptuais[1]. Como vimos acima, se a percepção é direta ou indireta não tem implicações significativas para a tese evolucionista. Tudo o que precisamos conceder é que as propriedades dos estados mentais são modos de apresentação de propriedades do ambiente, ou, para colocar de modo mais preciso, que elas constituem o acesso que temos ao mundo e que

[1] Sobre essa questão, ver Crane (2006) e mais recentemente Brogaard (2014) e Nudds (2009).

elas representam esse mundo de um certo modo.

Mas o que significa dizer que uma propriedade mental é capaz de representar propriedades do mundo? Defendendo uma versão da TR, Michael Tye (1995) sugere que um estado S representa uma propriedade P quando condições apropriadas são satisfeitas e S ocorre em um sistema A sse P e *porque* P. Isso quer dizer que S só pode representar P sse o ambiente E satisfaz certas condições como P ser a causa de S. Se esse for o caso, então podemos dizer que S representa P. Mas se A realiza S na ausência de P e fora de E, então A representa P de modo equivocado (*A misrepresents P*). Tye (1995), portanto, sustenta que a relação entre S e P é uma relação de *variação causal*, isto é, um estado representacional S está causalmente relacionado a P de modo que S é capaz de *rastrear* (*track*) a propriedade P: "A ideia central, portanto, é que uma representação diz respeito a uma covariação ou correlação (rastreamento [*tracking*], como comumente denomino esse processo) em situações adequadas". (TYE, 1995, p.101)

Embora Tye também defenda a TR, penso que não precisamos nos comprometer com essa concepção da relação entre representação e mundo. Mohan Matthen (2001) argumenta, por exemplo, que os $qualia$ das cores não rastreiam propriedades do mundo do mesmo modo que os $qualia$ de outras modalidades sensoriais. Matthen sustenta que não podemos treinar pessoas para discriminar propriedades mais básicas dos $qualia$ das cores do mesmo modo em que as treinamos no caso dos enólogos ou músicos profissionais. Se esse fosse o caso, então as propriedades representacionais dos $qualia$ das cores não precisam rastrear propriedades do mundo como os $qualia$ de outras modalidades o fazem[2].

A versão da TR de Dretske (1995) é capaz de acomodar esse caso. De acordo com Dretske, uma representação S não precisa ser uma representação verídica de algo presente no ambiente. Em outras palavras, S pode não rastrear nenhum aspecto real do mundo. Tudo o que é preciso para S ser uma representação é que ele tenha sido *designado* para representar algo:

> Um fato representacional S é um fato sobre o que S foi designado a fazer, um fato sobre a informação que ele deve trazer consigo. Existem fatos sobre representações — fatos sobre sua cor, forma, constituição material, e modo de operação — que não dizem nada sobre qual informação eles devem forne-

[2]Para mais sobre diferenças estruturais e qualitativas em modalidades diversas da percepção, ver O'Callaghan (2011, 2015)

> cer, ou, para ser mais preciso, se eles devem fornecer qualquer informação. (DRETSKE, 1995, p. 3)

Nesse sentido, se S representa ou não P de modo verídico diz respeito a se S foi designado a representar P veridicamente. De um ponto de vista evolutivo, isso quer dizer que S representa P se S foi realizado no passado e a realização de S ajudou os sistemas nos quais eles ocorreram a lidar com condições do ambiente de modo vantajoso frente aos sistemas que não realizavam S. Desse modo, podemos dizer que representações sensoriais nos dão informações relevantes sobre o mundo que era relevante para nossos ancestrais, o que os permitiram serem selecionados.

Assim, pode ser que haja algo como *sense-data* que sejam objetos diretos da experiência e que servem como intermediários para termos contato com os objetos do mundo. O problema em associar a TR com objetos como os *sense-data* é que não sabemos se os últimos possuem qualquer relação com o mundo, e em caso positivo, qual a natureza dessa relação. Esse é um problema filosófico genuíno, mas se estivermos dispostos a aceitar a tese minimalista de que a mente tem um papel importante na nossa sobrevivência, o que acredito ser uma tese bastante razoável, então essa questão não se coloca para nós. Isso se dá porque não faz sentido assumir que estados mentais não possam representar propriedades do mundo real de um ponto de vista evolutivo. Isso resultaria numa tese epifenomenalista radical, de acordo com a qual estados mentais não possuem eficiência causal.

É importante lembrar que a teoria epistêmica dos $qualia$ que defenderei aqui não se compromete com a existência de *sense-data*. A teoria epistêmica é, na verdade, neutra quanto à essa questão. O que pretendo mostrar, no entanto, é que se houvessem *sense-data*, poderíamos falar dessas entidades como veículos representacionais indiretos de alguma realidade externa, ainda que não pudessem rastrear causalmente o mundo, uma vez que um estado representacional, segundo a TR, depende do que ele foi designado a realizar.

Um último comentário é necessário aqui. Ambos Dretske (1995) e Tye (1995) negam a existência de *sense-data*, comprometendo-se com uma tese externalista em relação à natureza do conteúdo representacional das experiências conscientes. De modo mais específico, para eles "os $qualia$ não estão dentro da cabeça" (*qualia ain't in the head*). Um ponto importante

derivado dessa afirmação é que novas perspectivas de investigação surgem referente à questão de onde se localizam os $qualia$.

Gilbert Harman (1990), por exemplo, argumenta que nossas experiências perceptuais são *transparentes*. Isso quer dizer que quando experiencio um $quale$, não estou em contato direto — ou não experiencio diretamente — o $quale$ como constituinte da experiência, mas sim o $quale$ da coisa externa ao qual ele está relacionado. De modo resumido, para Harman um $quale$ é sempre representado como $quale$ em alguma coisa. O exemplo dado pelo autor é o de uma experiência perceptual de uma árvore. Quando olhamos para uma árvore, não temos acesso à experiência intrínseca das cores "marrom" e "verde". Temos, ao contrário, a experiência do marrom-na-árvore e do verde-na-árvore. Essa transparência dos $qualia$ parece indicar que essas propriedades não podem ser simplesmente abstraídas das experiências particulares.

Essa visão sobre a natureza da percepção contrasta com a afirmação de que existem *sense-data*. Se experiências são de fato transparentes e se não podemos experienciar a qualidade intrínseca à experiência, mas somente ao mundo, então parece não haver espaço para os *sense-data*, uma vez que não são necessários para explicar a natureza dos objetos da experiência. Mais ainda, o aspecto transparente da experiência parece sustentar a TR, já que os $qualia$ das experiências sensoriais não são algo intrínseco à experiência, mas algo que é representado *no* mundo, o que faz sentido do ponto de vista evolutivo.

Além disso, essas considerações nos permitem responder a uma questão bastante intrigante sobre a mente, isto é, se a mente pode ser reduzida ao cérebro — como supõem os reducionistas — então por que não encontramos os $qualia$ dentro de nossas cabeças? De acordo com Dretske (1995), isso ocorre porque os $qualia$ são representações, e como tais, eles são representados nas coisas e não no cérebro. Como o próprio Dretske (1995, p. 35) destaca, o cérebro é o *veículo representacional*, enquanto a mente é o *conteúdo* dessa representação.

Não pretendo com isso apresentar um argumento definitivo para descartar teorias *sense-data*, mas somente mostrar que a TR aliada a uma versão da tese externalista pode apresentar respostas para algumas questões difíceis da filosofia da mente. Mais importante ainda, meu objetivo é mostrar que TR mais o externalismo são teses plausíveis do ponto de vista evolutivo. A partir disso, podemos concluir que falar de estados mentais a partir de uma perspectiva evolutiva

exige considerá-los como entidades que representam propriedades do ambiente externo. Tendo isso em vista, podemos avançar para a discussão sobre uma explicação teleofuncionalista dos $qualia$.

Para facilitar a discussão, focaremos no caso das dores, uma vez que explicações teleofuncionais nesses casos é menos controversa. Dores parecem ter uma vantagem evolutiva bastante clara: elas nos alertam sobre situações em que algo não está certo com nosso corpo. Assim, se considerarmos dores como a representação de algo, podemos dizer que elas representam que algo não está certo em nosso corpo ou em parte dele. Isso nos alerta sobre possíveis perigos ou danos, permitindo assim a elaboração de estratégias para evitá-los. Parece pouco questionável também que dores são estados mentais e que possuem alguma função no organismos de seres vivos. Novamente, se procuramos uma explicação evolutiva de um estado mental, então é natural atribuir-lhes funções. Uma versão simplifcada da explicação funcionalista clássica pode ser dada nas seguintes linhas:

> **Dores:** são definidas pelo papel causal que exercem na mediação de $inputs\ I$ e $outputs\ O$ em relação aos estados mentais $m_1 \ldots m_n$ de um organismo E; sendo I = "danos nos tecidos" e O = "emissão de ruídos e estremecimentos".

Essa é a caracterização dada pelo funcionalista que aceita a noção matemática de função descrita por Lycan (1981). Para ver isso, note que a dor é definida a partir do seu papel de transição entre o domínio de $inputs\ I$ e o domínio de $outputs\ O$ aliado a sua relação com outros estados mentais de E.

Embora não sejam equivalentes, a explicação funcional tradicional permite identificar quatro variáveis que também aparecem na explicação teleofuncional: o item i, o sistema S, o comportamento C e a teoria T. O item "dores" tem uma função associada a partir da observação do comportamento "ruídos e estremecimentos" presentes em um sistema E (um ser humano, por exemplo) de acordo com teorias mais gerais da biologia T. Note que o propósito P e o ambiente A são deixados de fora da explicação tradicional.

Como vimos anteriormente, P é a variável mais importante de uma explicação teleofuncional, uma vez que a existência de um propósito é o que subscreve caráter teleológico às explicações funcionais. Além disso, P é compatível com a concepção etiológica de funções, o

que explicita também a importância do ambiente A. Para ver isso mais claramente, considere que a seleção natural seleciona traços que são benéficos de acordo com um ambiente específico. Se girafas com pescoços longos nascerem em um ambiente no qual a comida está localizada no chão, então suas chances de sobreviver — e consequentemente de se reproduzir — nesse ambiente serão drasticamente reduzidas[3]. Nessa situação, indivíduos com pescoços longos não seriam selecionados. A consequência disso é que fazer referência ao ambiente é essencial para as explicações evolutivas, visto que um traço pode ser avaliado somente no contexto do ambiente em que foi selecionado[4].

Podemos ver agora que as variáveis de Wimsatt (1972) são necessárias para uma explicação evolutiva dos estados mentais. A questão agora é explicar como elas diferem da explicação tradicional e como a introdução delas pode auxiliar na resolução dos problemas que discutimos. A resposta para a primeira questão é que explicações teleofuncionais diferem das explicações funcionalistas tradicionais porque são sensíveis a P e E. Essa diferença é o que explica a introdução da noção normativa de "mau-funcionamento" (*malfunction*). Isso é observado no caso de traços biológicos que possuem funções mas que nem sempre são capazes de realizar essa função de modo apropriado. Tendo isso claro, podemos ver agora como isso nos ajuda com problema epistêmico. Se os $qualia$ são considerados à luz da biologia evolutiva, então estados mentais como as dores podem ser descritos como um item i pertencente a um sistema S que foi selecionado para certos propósitos P a realizar o comportamento C em circunstâncias ambientais A. De modo mais preciso,

$i =$ estado mental (dor);
$S =$ qualquer sistema capaz de realizar esse estado (e.g., um ser humano);
$P =$ permitir ao sistema identificar algum dano ou perigo;
$B =$ gritos, estremecimentos e ações para aliviar a dor;
$E =$ danos nos tecidos ou órgãos em certas situações; e também
$T =$ teorias mais gerais da biologia.

Note que o ambiente pode ser definido também como as condições em que i realiza sua função. Na medida em que dores representam danos em S, então as condições ambientais

[3] Exploro esse ponto com mais detalhes em Sant'Anna (2014).

[4] Ver Darden e Cain (1989) para um argumento em favor dessa visão. Tratarei com mais detalhes dessa questão mais tarde nesse capítulo.

podem ser vistas como "danos no corpo em certas circunstâncias". Nesse contexto, o cálculo funcional é dado do seguinte modo:

> $F = i$ (dores) + S (seres humanos) + P (tornar S consciente de danos ou perigos) + C (gritos, estremecimentos e ações para aliviar a dor) + A (danos nos tecidos ou órgãos em certas circunstâncias) + T (teorias mais gerais da biologia) = evitar danos no corpo.

É importante notar que a caracterização dessas variáveis pode ser muito mais complexa do que a apresentada aqui. Ainda assim, isso não nos impede de explicitar a estrutura lógica de uma explicação teleofuncional das dores. Mais ainda, se A e P são satisfeitos, então podemos atribuir a F um caráter normativo. Podemos agora dizer que dores existem e que são realizadas porque foram selecionadas a realizar F.

Tendo essas considerações em vista, analisemos agora o caso do espectro invertido. O argumento, descrito a partir dos aspectos lógicos do teleofuncionalismo, pode ser colocado da seguinte forma: "Para um estado mental i, é possível inverter i sem mudar F". Poderíamos, por exemplo, colocar "cócegas" no lugar da "dor" e F se manteria intacto. Nesse cenário, o cálculo funcional seria dado da seguinte forma:

> $F = i$ (cócegas) + S (seres humanos) + P (tornar S consciente de danos ou perigos) + C (gritos, estremecimentos e ações para aliviar a dor) + A (danos nos tecidos ou órgãos em certas circunstâncias) + T (teorias mais gerais da biologia) = evitar danos no corpo.

De acordo com o teleofuncionalismo, essa formulação é equivocada. Para entender isso, note que ao inverter uma das variáveis de F, necessariamente o resultado final teria que ser $F_1 \neq F$ para $i_1 \neq i$. Isso leva à conclusão de que se cócegas são equivalentes a dores, tal que a inversão seja possível sem diferença funcional, então não há nenhuma diferença entre elas. Essa conclusão, no entanto, é pouco plausível do ponto de vista evolutivo, uma vez que os estados mentais com *qualia* não podem ser funcionalmente diferenciados, o que os impossibilita de ser objeto de seleção.

Um dualista substancial não veria problema com essa conclusão, já que desaloca os estados mentais para um domínio em que a seleção natural pode não operar. Mas se pretendemos evitar essa tese metafísica mais radical, então essa dificuldade tem que ser solucionada. Esse

problema não é, entretanto, o único problema que o espectro invertido tem frente ao teleofuncionalismo. Lembre-se da nossa discussão sobre a identificação heurística dos estados mentais com áreas do cérebro. Imagine agora que pudéssemos realizar essa identificação no caso das dores. Nesse contexto hipotético, a teoria *interfield* das dores seria bastante precisa. Teríamos um conhecimento preciso do que acontece no nível neuronal, assim como uma compreensão mais detalhada dos aspectos psicológicos e evolutivos das dores.

Considere agora a discussão sobre mau-funcionamento. Vimos que a seleção atribui normatividade à realização da função de um traço selecionado, o que nos permite dizer também que esse traço pode não realizar sua função apropriadamente, ou simplesmente mal-funcionar. A noção de mau-funcionamento requer, desse modo, que algo necessariamente se altere na realização da função de um traço. Nos casos biológicos, isso quer dizer que algo físico deve se alterar para que um traço ou órgão possa mal-funcionar.

Se pudermos dar uma explicação teleofuncional das dores, então as áreas cerebrais capazes de realizá-las devem ser capazes de mal-funcionar. Assim, se houver uma mudança em i como prescrita no caso do espectro invertido, então a função F necessariamente se altera, caso contrário violar-se-ia o caráter normativo das explicações teleofuncionais. Se F for a função das dores, e se F se altera, então dores estão mal-funcionando. Mas se dores são capazes de mal-funcionar, daí se segue que deve haver alguma diferença no substrato que a realiza. Mais ainda, se uma dor se altera de modo tão drástico e se torna cócegas, um teórico que possui uma teoria *interfield* das dores provavelmente conseguiria identificar essa mudança[5].

Além do já discutido, alguém poderia sustentar que o argumento apresentado esteja mal direcionado, uma vez que o que realmente define um estado mental com aspectos qualitativos são seus $qualia$. Embora não haja diferença funcional na explicação clássica, podemos identificar a inversão dos $qualia$ a partir do ponto de vista de primeira pessoa, uma vez que a experiência subjetiva das dores e das cócegas é diferente.

Acredito que essa objeção não oferece dificuldades ao teleofuncionalismo. Como destaquei anteriormente, essa objeção pressupõe a distinção teórica entre consciência de acesso e

[5]Note que se o investigador não observar a mudança, isso não se deve necessariamente a limitações lógicas do teleofuncionalismo, mas sim a limitações epistemológicas resultantes do conhecimento limitado do fenômeno que aquele investigador tem no momento.

consciência fenomenal. Argumentando contra essa separação, Cohen e Dennett (2011) defendem que os $qualia$ não constituem algo que existe em uma dimensão separada da constituição funcional do cérebro, o que implicaria, entre outras coisas, abandonar a distinção entre consciência de acesso e consciência fenomenal.

Para motivar essa tese, Cohen e Dennett formulam uma experiência de pensamento que eles chamam de "experiência perfeita" (COHEN e DENNETT, 2011, p. 361). Nesse experimento, os autores pedem que imaginemos um cenário ideal em que os cientistas tenham um mapeamento perfeito das funções cognitivas do cérebro. Isso os permitiria identificar as áreas responsáveis pelas experiências visuais, assim como delimitar com precisão áreas que processam informações sobre as cores, distância, forma, etc. Cohen e Dennett imaginam um caso no qual as capacidades cognitivas referentes ao processamento de informações sobre as cores sejam isoladas, de modo que um sujeito possa ver perfeitamente um objeto, mas não pode descrever a sua cor.

> Quando mostrado uma maçã colorida, o que os participantes do experimento hipotético diriam? Eles seguramente não diriam que veem qualquer cor, uma vez que as áreas responsáveis pelo processamento de cores foi isolada de outras áreas, inclusive aquela referente à produção de linguagem. Eles poderão identificar o objeto como uma maçã porque as áreas responsáveis pelos outros aspectos da cognição visual estão intactas e conectadas às outras regiões. Assim, eles se encontram num estado de cegueira de cores. Podemos imaginá-los dizendo "Eu sei que vocês dizem que minhas áreas de processamento de cores estão ativadas de um modo singular, e eu sei que vocês acreditam que eu tenho a experiência consciente de cores, mas eu olho para maçã, eu foco minha atenção nela, e ainda assim eu não tenho nenhuma experiência de cor. (COHEN e DENNET, 2011, p. 361)

Como Cohen e Dennett (2011, p. 362) enfatizam, qualquer tipo de consideração funcional sobre o cérebro elimina a hipótese de que esses indivíduos estejam vendo cores. O que justificaria, portanto, dizermos que ainda assim os sujeitos do experimento sejam conscientes, tal como o faríamos caso a distinção de Block (1980) for verdadeira? A partir dessas considerações, parece ser plausível concluir que há pelo menos algo de inconsistente em dizer que é possível ter uma experiência subjetiva sem correlação com aspectos funcionais.

Um caso similar ao das cores pode ser imaginado com as dores. Imagine que estejamos ainda nesse cenário de teorias bem desenvolvidas. Considere que alguém permita que um

cientista o atinja com uma faca. Se isolarmos as áreas cerebrais de modo que as áreas que processam informações sobre dores sejam separadas de todas as outras funções cognitivas, e depois pudéssemos mostrar ao sujeito seu braço machucado, ele diria algo do tipo "Eu sei que meu braço está sangrando e que minha pele está machucada, mas ainda assim não sinto nenhuma dor". Se isso estiver correto, então a distinção de Block (1980) parece ter pouco apelo nesses casos. Se continuarmos a sustentá-la, então temos que aceitar a conclusão pouco intuitiva de que os sujeitos da experiência perfeita veem cor — ou sentem dor — ainda que o sujeito relate não sentir nada.

A conclusão que quero explicitar é que a inversão de espectro não é possível quando definimos os $qualia$ teleofuncionalmente. Os casos que consideramos, entretanto, são casos em que a inversão ocorre *depois* que o sujeito nasceu. Nesses casos, o conjunto completo de informações sobre o cérebro dos sujeitos do experimento exigiria a comparação de informações de *antes e depois* da inversão. Isso gera um problema na medida em que é logicamente possível pensar no caso de pessoas que já tenham nascido com seus espectros invertidos. Essa também é uma questão que o teleofuncionalismo tem que responder se pretende superar o problema epistêmico.

No capítulo 3, discutimos a noção de função etiológia, que Ruth Millikan (1984) chama de funções próprias. Uma concepção importante desenvolvida por Millikan em relação às funções etiológicas é a noção de *reprodução*. De acordo com Millikan (1984), existem dois modos pelos quais podemos dizer que B é uma reprodução de A. O primeiro é ilustrado pelo caso do papagaio que diz "oi" porque ouviu alguém dizendo "oi". Nesse caso, o papagaio reproduz de modo direto aquilo que ouviu. A palavra "oi" é uma cópia direta daquela emitida por algum falante da língua portuguesa. De acordo com Millikan, dizemos que a ocorrência da palavra "oi" emitida pelo papagaio é um membro de uma *família de reprodução estabelecida de primeira ordem* (*first-order reproductively established family*).

O segundo caso mencionado por Millikan é ilustrado por casos biológicos. Considere os corações. O coração do filho de John é certamente uma reprodução do tipo "coração humano", mas ele não é uma reprodução *direta* do coração de John. Ele é uma reprodução dos genes de John que carregam a informação genética que torna possível o desenvolvimento de corações

humanos. Nesse sentido, se John desenvolver uma doença do coração que não seja causada por fatores genéticos, seu filho não nascerá com essa doença. Isso ocorre porque o coração do filho de John é apenas uma reprodução *indireta* do coração de John. Millikan diz que traços ou órgãos biológicos como corações são membros de uma *família de reprodução estabelecida de ordem superior* (FOS) (*higher-order reproductively established family*).

Consideremos essas noções com mais detalhes, visto que serão de suma importância na discussão que se segue. Como vimos, traços biológicos são resultados de um processo seletivo. Como tal, meu coração não é uma reprodução direta do coração de meus pais, mas sim do coração que foi objeto de seleção no passado. Assim, meu coração é membro de uma FOS. Se depois de nascer, eu desenvolver uma doença do coração, o médico terá como parâmetro de diagnóstico o funcionamento do tipo selecionado no passado, e não propriamente o modo em que meu coração funcionava antes da doença se desenvolver.

De modo similar, as mesmas considerações são aplicadas ao caso dos $qualia$. O estado mental de dor que tenho quando parte do meu corpo se machuca não é uma reprodução direta do estado mental que meus pais têm nessas ocasiões. Pode ser o caso, por exemplo, que meus pais tenham sofrido algum dano neurológico de modo que não sintam mais dores, e mesmo que eu tenha nascido depois desse acidente, isso não significa que eu terei os mesmo problemas neurológicos. Isso ocorre porque estados mentais são membros de uma FOS, o que os torna reproduções de estados mentais selecionados no passado, e não de estados mentais de nossos pais ou avós.

Essas considerações são bem sugestivas para responder ao problema da inversão antes do nascimento. De acordo com o teleofuncionalismo, se uma pessoa nascer com o espectro invertido, então necessariamente há uma diferença funcional, caso contrário as dificuldades destacadas no caso da inversão depois do nascimento também se aplicam. O que torna a inversão impossível é o fato de estados mentais serem membros de uma FOS, e os membros dos quais eles são reproduções não podem ser invertidos. Nesse sentido, as considerações sobre normatividade também se aplicam nesse caso. Se uma pessoa nasce com o espectro invertido, então é preciso que haja diferença funcional, caso contrário os $qualia$ perdem seu valor evolutivo.

Podemos agora considerar o que isso tudo nos diz sobre o problema epistêmico. A con-

clusão geral é a de que o problema epistêmico é resolvido pela concepção etiológica de função, mais especificamente, pelo seu caráter normativo. Desse modo, temos agora uma definição de função e do funcionalismo que permite a justificação lógica da nossa crença de que outros seres humanos possuem $qualia$ (pelo menos nos casos específicos discutidos até agora). Superamos assim a dificuldade associada à contingência de um raciocínio indutivo.

De um modo mais específico, ainda que $qualia$ e estrutura funcional não sejam logicamente relacionados, posso dizer com segurança que os meus $qualia$ estão associados a minha estrutura funcional. Posso dizer também que, como ser humano, sou resultado de um processo longo de evolução. Sei também que outros seres humanos e outros animais são resultados desse processo seletivo, e que, portanto, compartilhamos dos mesmos ancestrais. Agora, se mentes são reproduções de mentes ancestrais, e se meus estados mentais são membros de uma família de reprodução de ordem superior, então as mentes de todos os seres humanos são reproduções indiretas de um modelo de mente ancestral. Isso quer dizer que o $quale$ que tenho ao olhar para uma rosa é o mesmo $quale$ que você e outros seres humanos possuem quando olham para essa rosa. Isso se dá porque nossos $qualia$ são reproduções de um mesmo $quale$ ancestral. Mais ainda, de acordo com a noção normativa de função, se alguma diferença ocorrer, então necessariamente haverá uma diferença funcional. Nesse sentido, podemos justificar nossa crença sobre os $qualia$ de outros seres humanos, e também formular a primeira parte da *teoria epistêmica dos qualia*:

> (TEQ_1) Podemos justificar nossas crenças que outros seres humanos possuem os mesmos $qualia$ que nós porque esses $qualia$ são membros de uma FOS (família de reprodução de ordem superior), e como tal, são reproduções de um mesmo $quale$ ancestral. Se houver uma diferença nos $qualia$, então deve haver também uma diferença na funcionalidade do material substrato, uma vez que o que tem uma função também deve ser capaz de mal-funcionar.

A TEQ_1 se restringe somente a casos biológicos. Ela resolve o problema epistêmico somente no caso dos seres humanos e animais. O funcionalismo, no entanto, não elimina a possibilidade de que sistemas não-orgânicos também tenham mente. Essa é uma das principais consequências da tese da múltipla realização. O problema é que a TEQ_1 não é capaz de resolver o problema epistêmico para sistemas que não são objetos de seleção natural, o que exige diferentes considerações que abordaremos na próxima seção.

4.2 Aspectos gerais da seleção

A discussão apresentada até aqui só é capaz de oferecer uma solução para os casos biológicos, isto é, a TEQ_1 só se aplica aos sistemas que são objetos de seleção natural. Como vimos na capítulo 3, entretanto, tanto traços e órgãos biológicos quanto artefatos parecem ser objetos de processos de seleção. Mais ainda, o funcionalismo, pelo menos em sua formulação tradicional, parece aceitar que o termo função se aplica inequivocamente a ambos os domínios. A questão que se coloca agora é se o mesmo pode ser dito do teleofuncionalismo. De modo mais específico, como conciliar a concepção de função etiológica com a tese da múltipla realização (MR)?

De acordo com a MR, o substrato em que a mente se realiza não é importante. O que importa, em última instância, é o papel funcional de um estado mental dentro de um sistema, seja qual for o substrato em que é realizado. A tese da MR é de particular importância para aqueles que acreditam que mentes se estendem a sistemas não-orgânicos. Assim, se a MR for plausível, então o funcionalismo é capaz de abarcar teoricamente casos em que robôs ou computadores tenham mente.

Tendo isso em vista, a questão que nos interessa nessa seção é a seguinte: seria possível construir uma versão da TEQ para casos de sistemas não-orgânicos que não são resultados de seleção natural? Como uma primeira aproximação, pode-se argumentar que a TEQ_1 se aplica ao caso de sistemas não-orgânicos porque ambos traços biológicos e artefatos são resultados de processos de seleção. Embora seja atrativa, essa estratégia é vista com suspeitas por alguns filósofos.

Em seu artigo "*Functions as selected effects: the conceptual analyst's defense*", Karen Neander argumenta que essa aproximação não é possível porque a seleção natural e a seleção artificial (ou intencional) são processos essencialmente distintos. Neander apresenta três argumentos para defender essa afirmação. Primeiro, ela argumenta que funções próprias — no sentido de Millikan (1984) — em casos biológicos se aplicam primariamente a tipos e apenas secundariamente a ocorrências (Neander, 1991b, p. 174). Isso quer dizer que, para Neander, podemos falar apenas de funções próprias quando consideramos o objeto de explicação enquanto uma classe geral, isto é, como um tipo. A função atribuída à ocorrência de um coração, por

exemplo, seria de natureza derivada e secundária, uma vez que os objetos de seleção são tipos, e não ocorrências. Isso ocorre na seleção natural, de acordo com Neander, porque ela não seleciona indivíduos, mas sim variações que ocorrem entre gerações de indivíduos que são dotados de um determinado traço vantajoso em termos de sobrevivência. Para observar a diferença, note que no caso da seleção artificial as atribuições funcionais parecem ser feitas a um protótipo particular, e não propriamente a um tipo. Designers, por exemplo, não precisam criar necessariamente diversos modelos de garfos para saber, ou ao menos ter uma ideia, de qual modelo será mais apropriado. Nesse sentido, não há variações de indivíduos através de gerações, o que parece sustentar o argumento de Neander.

A segunda dificuldade levantada por Neander (1991b, p. 174) refere-se ao fato de que traços biológicos são selecionados por causa da função que realizaram no passado, enquanto artefatos são selecionados por aquilo que fazem no presente, ou pelo que farão no futuro. Pelo fato de a seleção natural não ser intencional, não é possível que um traço seja selecionado pela sua função futura, uma vez que o processo seletivo não pode saber *a priori* as vantagens daquele traço. Isso indica, novamente, uma assimetria entre ambos os processos, uma vez que um engenheiro não precisa necessariamente realizar inúmeros experimentos com variações individuais.

O terceiro argumento de Neander (1991b, p. 175) está conectado ao segundo argumento, uma vez que diz respeito à performance de um traço no processo de seleção. Em casos biológicos, como mostra o segundo argumento, a função de um traço deve ser realizada no passado. Já no caso de artefatos, quando engenheiros criam um projeto de freios automotivos mais eficientes, eles não precisam construir vários modelos possíveis e testar cada um em situações específicas. O que eles fazem, ao contrário, é criar um modelo inicial baseado em uma seleção prévia feita num nível abstrato, e depois testam apenas alguns modelos. Nesse sentido, um freio não precisa, necessariamente, realizar sua função no passado para ser selecionado.

As objeções de Neander colocam sérios problemas para a teoria epistêmica, uma vez que parece tornar impossível a conciliação do teleofuncionalismo com a tese da MR. O problema com essa conclusão é que não podemos eliminar de antemão a possibilidade de sistemas não-orgânicos terem mente. Assim, o teleofuncionalismo deve oferecer uma resposta satisfatória a

essas objeções.

Acredito que uma solução é possível, e tentarei mostrar isso agora. Em seu artigo *Function and Design*, Philip Kitcher (1993) argumenta que ambos os processos de seleção natural e seleção artificial podem ser aproximados através do conceito de $design$. De acordo com Kitcher, explicações funcionais estabelecem uma ligação direta entre o design associado a um traço biológico e a seleção natural, de tal modo que sentenças como "A função de X é o que X foi designado a fazer, e o que X foi designado a fazer é o motivo pelo qual X foi selecionado" (KITCHER, 1993, p. 263) são capazes de expressar o estatuto funcional de um certo item ou traço biológico.

A relação entre design e seleção é um elemento bastante comum em discussões que antecedem os trabalhos de Darwin. Em outras palavras, a noção de design parece implicar a existência de um agente intencional inteligente capaz de criar coisas com propósitos determinados. Como o nome explicita, esse designer inteligente pode prever os elementos essenciais para a construção de um traço ou artefato, tendo em vista o objetivo de sua criação.

Darwin nos mostrou, no entanto, que postular a existência de um designer inteligente não é condição necessária para explicar o design de entidades biológicas. Para justificar essa tese, Darwin argumenta que há um processo de seleção que opera de modo lento e mecânico durante um largo período de tempo, o que possibilita aos organismos complexos ter sua origem explicada a partir de organismos mais simples. De modo resumido, como aponta Dennett (1995), Darwin mostrou que com variáveis como tempo e caos é possível obter organização e design. O que é importante notar é que o tipo de design associado à seleção natural, ainda que não esteja associado à ideia de um agente intencional, parece bem similar ao tipo de design presente na seleção artificial. Assim, uma aproximação entre ambos os processos parece ser promissora. Mas como isso é possível?

Uma resposta pode ser encontrada no trabalho de Darden e Cain (1989). Nesse artigo, Darden e Cain explicitam os aspectos lógicos e abstratos da seleção natural, e posteriormente comparam esse modelo lógico com outros processos de seleção que ocorrem na ciência. De modo resumido, a proposta dos autores é observar se os aspectos lógicos da seleção natural aparecem em outros processo de seleção da ciência. Vejamos isso com mais detalhe.

Os casos de seleção que Darden e Cain examinam são os de seleção para formação de anticorpos e a seleção para funções cerebrais. Não tratarei desses detalhes, dando enfoque somente à discussão lógica subjacente ao trabalho dos autores. O que nos importa, de modo particular, é o modelo abstrato, ou a *teoria-tipo* (*type theory*), da seleção natural explicitado por Darden e Cain. A minha estratégia será discutir os aspectos desse modelo abstrato e ver se ele se aplica ao caso da seleção artificial. Se esse for o caso, então podemos visualizar uma solução para os problemas de Neander (1991b).

Darden e Cain apresentam cinco aspectos lógicos importantes da seleção natural. O primeiro se refere à necessidade de explicitar as pré-condições (aspecto A) exigidas para haver seleção natural. Como argumenta Neander (1991b), para haver seleção natural é preciso que haja variações de ocorrências em um ambiente comum. Como essas ocorrências se encontram em um ambiente comum, elas também interagem causalmente (aspecto B) com esse ambiente. Esse último ponto é particularmente importante porque permite definir a adaptabilidade de uma ocorrência no ambiente.

> Analisamos o ambiente em termos do fator crítico que faz uma certa propriedade ser causalmente relevante para a interação seletiva. O fator ambiental crítico e a interação da propriedade variante com o ambiente são fundamentais para o efeito subsequente da interação seletiva. (DARDEN; CAIN, 1989, pp. 190-1)

A importância do efeito de um traço na interação com o ambiente nos leva ao terceiro aspecto, isto é, os efeitos (aspecto C). De acordo com esse aspecto, a seleção natural deve selecionar algo para algum efeito específico. Funções básicas para a sobrevivência, como obter comida ou fugir de predadores, são exemplos simples de efeitos.

Por fim, os efeitos apontam para os dois últimos aspectos apresentados por Darden e Cain (1989, p. 191). Esses aspectos são os de que a seleção natural deve permitir a presença de efeitos de longo alcance (aspecto D) e efeitos de mais longo alcance ainda (aspecto E). A principal ideia por trás de D e E é que a seleção natural deve permitir o aumento das taxas de reprodução de um sistema S que possui um traço selecionado (aspecto D). "E" é, desse modo, uma extensão de D, na medida em que D pode também apresentar efeitos de longo alcance para S.

Darden e Cain (1989) resumem essa discussão através do seguinte esquema:

> A. Pré-condições
>
> i. Um grupo de Ys existe;
> ii. Ys variam no que diz respeito a possuírem uma propriedade P ou não;
> iii. Ys se encontram em um ambiente A com um fator crítico F.
>
> B. Interação
>
> iv. Ys, em virtude de possuírem ou não possuírem P, interagem de modo diferente com A;
> v. O fator crítico F afeta a interação, de modo que
>
> C. Efeito
>
> vi. ter P faz que Ys com P sejam beneficiados e Ys que não tenham P prejudicados;
>
> D. Efeitos de longo alcance
>
> vii. C deve ser acompanhado de um aumento de reprodução de Ys com P ou um aumento de reprodução com algo associado com Ys.
>
> D. Efeitos de mais longo alcance
>
> viii. D deve ser acompanhado de benefícios de longo alcance. (DARDEN; CAIN, 1989, pp. 192-3)

Darden e Cain consideram outros processos de seleção frente a esse esquema teórico. Quando esses casos se encaixam logicamente no esquema, Darden e Cain afirmam que eles implementam o mesmo tipo abstrato de teoria que a seleção natural apresenta. No que se segue, farei o mesmo que Darden e Cain, mas considerarei o caso particular da seleção intencional.

Para iniciar a discussão, proponho a retomada das variáveis lógicas das explicações teleofuncionais. Como vimos, essas variáveis eram amplas no sentido de não estarem restritas a casos biológicos. O primeiro passo, portanto, é ver se explicações teleofuncionais são possíveis para casos de artefatos. Consideremos um caso particular de um artefato como um refrigerador. Para que um refrigerador possa realizar sua função, cada item particular deve ter um termostato que permita regular sua temperatura. Desse modo, podemos construir a explicação teleofuncional do seguinte modo:

$i =$ termostato;
$S =$ refrigerador;
$P =$ controlar e medir a temperatura do refrigerador;
$B =$ mandar sinal x quando a temperatura estiver mais alta do que o normal e mandar sinal y quando a temperatura for mais baixa do que o normal;
$E =$ ambientes em que a temperatura varia de 0ºC e 50ºC;
$T =$ teorias mais gerais da termodinâmica e da física.

O que nos dá a seguinte fórmula do cálculo funcional:

$F = i$ (termostato) $+$ S (refrigerador) $+$ P (medir temperatura interna de S) $+$ B (mandar sinais específicos quando variações de temperatura ocorrerem) $+$ E (ambientes no qual a temperatura varia de 0ºC a 50ºC) $+$ T (teorias gerais da física e da termodinâmica) $=$ controlar a temperatura interna.

É possível notar, portanto, que uma explicação teleofuncional dos termostatos é possível. Gostaria de enfatizar o aspecto P novamente. Note que engenheiros escolhem o termostato pelo propósito que ele tem no funcionamento do refrigerador. A realização do propósito permite ao termostato realizar sua função, contribuindo assim para o funcionamento do refrigerador.

Considere agora que não existe apenas um tipo de termostato no mercado, mas sim uma grande variedade deles que são mais ou menos úteis de acordo com as necessidades de um projeto. Voltando às condições estabelecidas por Darden e Cain, podemos dizer que essa variação satisfaz o aspecto A, uma vez que existem diferentes itens i que diferem entre si.

A. Precondições

i. Um conjunto de termostatos existe;

ii. Termostatos variam de acordo com seu propósito P;

iii. Termostatos estão em um ambiente A com um fator crítico F (temperatura entre 0ºC e 50ºC)

Note que o item (ii) se refere à questão de se os termostatos são mais ou menos aptos a realizar seus propósitos. Isso exige um esclarecimento. Imagine que usemos um termostato industrial em refrigeradores comuns. Embora seu propósito seja o de controlar a temperatura, eles não realizarão sua função adequadamente porque a temperatura que podem medir é muito

alta. Nesse sentido, não seriam capazes de auxiliar no funcionamento do refrigerador. Não há, desse modo, variação de P, mas há variação no que diz respeito a i realizar seu propósito ou não, o que consiste em um caso similar, uma vez que se há uma propriedade P ou capacidade de realizar P, então S possivelmente será selecionado. Do mesmo modo, se P não existe ou se S não pode realizar P, então S provavelmente não será selecionado. Se isso estiver correto, então a seleção artificial ou intencional satisfaz o aspecto A.

Olhemos agora para B (interação). Imagine que construíssemos dois refrigeradores idênticos R_a e R_b, mas em um deles colocamos um termostato adequado "a" e no outro um termostato industrial "b". O aspecto da interação diz que itens i devem interagir com o ambiente A para serem selecionados. Claramente, nesse caso ambos os termostatos interagem com o ambiente, mas somente R_a realiza seu propósito. Podemos, portanto, caracterizar B:

B. Interação

iv. Termostatos, por realizarem ou não P, interagem de modo diferente com o ambiente A; e

v. fatores críticos F (temperatura entre 0ºC e 50ºC) afetam a interação de modo que...

Note também que o fator crítico F é diretamente responsável por como as interações entre R_a, R_b e o ambiente A acontecem. Se a temperatura fosse algo em torno de 2000ºC, então R_b serviria seu propósito enquanto R_a seria inútil. Desse modo, podemos dizer que a seleção artificial ou intencional também satisfaz B.

Já em relação a C, a questão diz respeito aos efeitos que R_a e R_b terão em A. Como vimos, R_a e R_b terão diferentes efeitos em A que serão benéficos ou não, o que consequentemente os tornarão mais adequados ou não a A. Assim, podemos dizer que a seleção artifical ou intencional satisfaz C:

C. Efeito

vi. realizar P faz com que os termostatos que realizam P sejam beneficiados e os que não realizam sejam prejudicados

Tendo esclarecido esses aspectos, sobram D e E. Esses dois podem ser tratados juntos. D diz que os efeitos que R_a tem em E devem ser seguidos de um aumento na reprodução de termostatos e E diz que esse aumento deve ter benefícios de longo alcance. Para ver como isso é possivel na seleção artificial, note que se R_a realizar seu propósito em E e E satisfaz os padrões de temperatura da Terra, então R_a venderá muitas unidades. O sucesso na venda criará a demanda por mais termostatos do tipo "a", o que fará com que sua produção seja aumentada, já que R_a realizará sua função corretamente e "sobreviverá" no mercado. Nesse caso, D é satisfeito. Note também que a sobrevivência de R_a no mercado pode torná-lo mais popular, o que tornará sua demanda maior pelos consumidores. Nesse sentido, E também é satisfeito, uma vez que o aumento de popularidade aumenta a produção de R_a.

D. Efeitos de Longo Alcance

vii. C pode ser acompanhado por um aumento na reprodução de termostatos que realizam P ou que reproduzem algo associado a termostatos

E. Efeitos de mais Longe Alcance

viii. D pode ser acompanhado por benefícios de maior alcance (popularidade no mercado e aumento de vendas)

Se essas considerações estiverem corretas, então podemos dizer que a seleção artificial ou intencional se encaixa na teoria-tipo da seleção natural. Isso justifica nossas intuições de que ambas eram bastante similares. Além disso, também suporta a afirmação de Kitcher (1992) que se destacou em nossa discussão. Em outras palavras, quando dizemos que algum objeto foi designado a algo, dizemos, no final das contas, que o objeto foi criado com um certo propósito.

A nossa preocupação agora se volta para as objeções de Neander (1991b). Agora que sabemos que seleção natural e seleção artificial pertencem a mesma teoria-tipo, podemos enfrentá-las com melhores recursos. As objeções colocadas por Neander parecem estar preocupadas com a ausência de um processo de tentativas e erros que é característico da seleção natural, mas que não se encontra na seleção artificial. Um aspecto importante que permite a seleção

natural explicar a complexidade dos sistemas orgânicos sem apelar para a existência de um ser criador inteligente é o seu caráter de operação por tentativa e erro durante espaçosos períodos de tempo.

Note que essa preocupação sustenta a primeira objeção. Quando Neander diz que funções próprias são atribuídas apenas a tipos, ela quer dizer que não podemos falar de uma única ocorrência como sendo selecionada para realizar uma função própria. Falamos, ao contrário, de várias ocorrências que foram selecionadas a realizar essa função. Nesse sentido, não é uma ocorrência individual que é selecionada, mas sim o tipo do qual ela faz parte. Isso quer dizer que a seleção natural não é um processo que ocorre rapidamente, no sentido de ser realizada em uma única geração de indivíduos. A seleção natural é, ao contrário, um processo mecânico que opera durante várias gerações. Isso implica dizer que ocorrências não podem ser selecionadas, uma vez que são sempre restritas a uma única geração. Torna-se claro, portanto, que se um processo seletivo operasse em uma ocorrência, ele violaria o requisito do processo de tentativa e erro.

O primeiro argumento de Neander não oferece um cenário incontornável para a analogia que proponho. Não podemos negar que a seleção artificial não envolve um processo de tentativa e erro semelhante ao da seleção natural, mas isso é uma limitação de caráter prático. Em outra palavras, embora não ocorra praticamente, isso não quer dizer que a seleção artificial não pressuponha o processo de tentativa e erro logicamente. O fato de os engenheiros não criarem diferentes individuais que variam de acordo com certo traço T não quer dizer que eles não tenham considerado essas variações. Na verdade, seria pouco prático (e economicamente inviável) criar refrigeradores com todos os termostatos possíveis. Os engenheiros podem prever, de antemão, qual desses termostatos funcionariam melhor em certas situações, e, depois disso, podem selecionar os mais adequados para os propósitos do refrigerador. No caso da seleção artificial, portanto, não há um processo de erro e tentativas concreto, mas ele está pressuposto, de maneira lógica, no produto final.

O argumento oferecido aqui, portanto, enfatiza que a diferença entre seleção natural e seleção artificial reside no âmbito prático, isto é, em como são realizadas concretamente. Isso se torna evidente, em parte, pelo fato de que ambos os processos pertencem a mesma teoria-

tipo. Para ilustrar esse ponto através de um exemplo, imagine que pudéssemos criar vários refrigeradores e então esperar para ver qual deles sobreviveria no mercado. Imagine ainda que pudéssemos realizar esse processo por várias gerações. Poderíamos dizer, ainda assim, que ambos os processo são diferentes, ainda que os resultados finais sejam os mesmos?

Mas o que tudo isso nos diz sobre a questão da seleção de tipos e não de ocorrências? De acordo com o argumento defendido aqui, a distinção de Neander é apenas aparente, uma vez que confunde considerações dos domínios práticos e teóricos. Em outras palavras, embora pareça ser o caso de um ponto de vista prático, o item que o engenheiro seleciona não é uma ocorrência frente a outras ocorrência, mas sim um tipo que foi considerado entre outros tipos num contexto hipotético.

Acredito que uma resposta a segunda e terceira objeções podem ser formuladas de modo similar. Iniciando com a segunda, o fato de que artefatos não são selecionados pela função que tiveram no passado é apenas uma diferença contingente dos processos de seleção. Para ver isso mais claramente, considere o cenário em que criamos vários refrigeradores com variações de termostato. Termostatos que forem similares ao tipo de termostato comum de refrigeradores serão selecionados frente a termostatos similares a um industrial, por exemplo. Desses termostatos selecionados, aqueles que são ainda mais similares serão selecionados, e assim sucessivamente. Nesse caso, o refrigerador realizou sua função antes de ser selecionado. O que esse cenário nos mostra é que não há mudança no produto final uma vez que o processo seletivo continua sendo essencialmente o mesmo. Novamente, o que muda é como o processo é realizado e não sua natureza.

Se o argumento contra a segunda objeção estiver correto, então a terceira objeção também não deve nos preocupar, uma vez que a realização da função na seleção natural e a ausência dessa realização na seleção artificial é apenas uma diferença prática. Note que ambas segunda e terceira objeções também estão fundamentadas na ausência de um processo de tentativa e erro. Isso ocorre porque a seleção de uma função em termos de sua realização no passado é algo característico do processo de tentativa e erro, uma vez que a seleção natural não pode prever qual o melhor tipo a ser selecionado. No caso da seleção artificial, essa função não precisa ser realizada porque designers inteligentes podem prever as consequências da realização de dife-

rentes tipos, o que nos permite dizer que as funções foram realizadas em um certo sentido (na mente do engenheiro, por exemplo), embora não tenham sido realizadas praticamente.

Para concluir essa discussão, podemos resumir a resposta a Neander (1991b) do seguinte modo: as diferenças apontadas por Neander são apenas contingentes, isto é, não destacam diferenças nos aspectos essenciais dos processos de seleção, mas somente no modo em que ambos são realizados. Nesse sentido, tendo em vista que seleção natural e seleção artificial pertencem a uma mesma teoria-tipo, podemos afirmar, sem nos preocupar com as objeções de Neander, que ambas são processos da mesma natureza. Essa última afirmação é de extrema importância para a teoria epistêmica em relação à tese da múltipla realização dos estados mentais.

4.3 Múltipla realização, $qualia$ ausentes e teleofuncionalismo

A discussão sobre a relação da seleção natural e da seleção artificial ou intencional foi guiada pela assunção de que se sistemas não-orgânicos podem ter, em princípio, mentes, então devemos explicar como a teoria epistêmica se aplica a esses sistemas. Vimos que o processo seletivo é o aspecto central da teoria epistêmica nos casos biológicos. Explicações baseadas na seleção tem um aspecto teleológico, e esse aspecto teleológico é o que fundamenta a teoria epistêmica.

Argumentei que a teoria epistêmica precisa do teleofuncionalismo para ter sucesso. Nos capítulos 1 e 2, discuti o funcionalismo com mais detalhes. Uma das teses que o permite superar dificuldades colocadas ao behaviorismo e à teoria da identidade é a múltipla realização dos estados mentais. Tendo isso em vista, o meu objetivo nessa seção é mostrar como o teleofuncionalismo pretende lidar com a tese da múltipla realização.

De modo mais preciso, minha sugestão é que a aproximação proposta entre processos de seleção naturais e artificiais pode finalmente responder esse problema. Como vimos, se quisermos estudar os $qualia$ a partir de uma perspectiva evolutiva, então temos que conceder que eles possuem a propriedade de representar aspectos do ambiente, uma vez que, como nos mostram Darden e Cain (1989), não pode haver adaptação de traços biológicos sem a interação com o ambiente. Nesse sentido, o meu ponto de partida aqui é o de que estados mentais com $qualia$ representam aspectos do ambiente.

Outro aspecto importante a ser lembrado é que os $qualia$ precisam ter uma função para

terem qualquer valor evolutivo, uma vez que sem ter esse aspecto funcional, não podem ser objeto de seleção. Desse modo, parece razoável dizer que os $qualia$ são selecionados porque sua função é representar aspectos do ambiente. Essa assunção, no entanto, traz como consequência o fato de que o define um $qualia$ de uma espécie S é a relação funcional e representacional entre um estado mental *m* de um sistema *s* e algum aspecto do ambiente *p*.

Se aceitarmos ambas as teses acima, então torna-se claro que o que define a identidade de um $qualia$ é sua função, e o que pode nos dizer se indivíduos da espécie S possuem o mesmo $qualia$ quando possuem um estado mental *m* é se esse estado mental foi selecionado para representar uma propriedade *p* do ambiente. Desse modo, podemos estabelecer o seguinte critério de identidade:

> (ID) Para todos os indivíduos $s_1 \ldots s_n$ da espécie *S* em relação com um ambiente *E* com propriedades $p_1 \ldots p_n$, se os membros de S estiverem em uma relação adequada *r* com E, tal que s_n instancia um estado mental *m* que representa p_n, então para todo *s* que se encontrar em *r* com *p*, *m* será o caso.

Uma questão natural é perguntar o que determina que s_1 e s_2, quando em uma relação r com p, terão o mesmo estado mental m com $qualia$? O que define isso é o fato de que o estado mental m foi selecionado para aquela função, e uma vez que a seleção é de caráter teleológico, isso permite estabelecer um critério *normativo* para a relação. Nesse ponto, a conciliação da tese da múltipla realização com o teleofuncionalismo se torna um pouco mais clara, isto é, se a seleção é o que determina que indivíduos $s_1 \ldots s_n$ pertencentes a S têm um estado mental m quando em relação r com uma propriedade p do ambiente E, e se a seleção natural e a seleção artificial são processos seletivos da mesma natureza, então se robôs $r_1 \ldots r_n$ forem selecionados para representar certa propriedade p do ambiente E quando estiverem em uma relação r com p, então, do caráter normativo dos processos seletivos, podemos dizer que os robôs $r_1 \ldots r_n$ terão o mesmo estado mental m que indivíduos biológicos membros de S possuem (assumindo que os membros de S e R têm a mesma história seletiva no que se refere a E).

De modo mais formal, a teoria de Millikan (1984) se torna essencial para formular esse argumento. Considere o fato de que um estado mental com $qualia$ seja membro de uma *família de reprodução de ordem superior* (FOS). Considere o caso das dores. Todas as instanciações de dores pertencem a uma FOS do tipo P. Isso quer dizer que todas instanciações p_n de dores

serão reproduções de um tipo que foi selecionado para certa função.

Antes de proceder, consideremos novamente o que Millikan (1984) entende por *reprodução*, mas dessa vez de um ponto de vista mais formal. Para Millikan, um indivíduo B é a reprodução de um indivíduo A quando:

> (i) B é similar a A em algum aspecto;
> (ii) essa similaridade é explicada por uma *lei in situ*; e
> (iii) as propriedades de reprodução estabelecidas $q_1 \ldots q_n$ que A e B têm em comum devem ser explicadas por uma *lei in situ* que correlaciona determinados sob um determinante.

Analisemos essas afirmações com mais detalhes. Considere o exemplo de uma máquina copiadora dado por Millikan (1984, p. 20). Imagine que precisemos fazer uma cópia de um texto de uma página. Esse texto, por algum motivo desconhecido, está escrito em três cores diferentes. O primeiro terço dele está em vermelho, o segundo terço está em verde e o resto está em azul. Ao notarmos que esse texto tem três cores, e não sabendo qual o seu objetivo, decidimos que o mais adequado seja fazer uma cópia fiel dele. Quando ligamos a máquina para realizar uma cópia, algumas leis da física operam sobre o funcionamento interno da máquina (as leis do eletromagnetismo, por exemplo). Se focarmos nossa análise em um nível mais geral, poderíamos descrever esses processos físicos de um modo um pouco mais familiar, dizendo, por exemplo, qual é a função da máquina, o que ela foi designada a fazer, etc. Poderíamos dizer que a função (no sentido de Cummins) de certa parte da máquina é a de reconhecer ocorrências de letras e cores. De modo simples, podemos dizer que essa parte faz algo do tipo: "Se a cor de uma ocorrência x é y (em que x é a letra reconhecida e y é a cor da letra), então imprimir ocorrência x_1 com cor y_1 (em que x_1 e y_1 são reproduções de x e y respectivamente)". Esse condicional descreve de modo mais ou menos preciso os processos físicos responsáveis pela cópia das ocorrências de letras, estabelecendo assim uma relação de regularidade causal entre o documento copiado (A) e a cópia do documento (B). Isso é o que Millikan chama de *lei in situ*.

Dada essas considerações, podemos agora analisar se a cópia do documento satisfaz as três condições acima. Dado que B é uma cópia perfeita de A, (a) é instantaneamente satisfeito. Além disso, considerando que a máquina funciona do modo descrito acima, temos uma *lei in situ* que explica por que B tem as propriedades $p_1 \ldots p_n$ em comum com A. Desse modo, (b)

também é satisfeito. Mas o que dizer de (c)? O que seriam determinantes e determinados? Millikan (1984) define esses termos do seguinte modo:

> Uma propriedade é "determinada" em relação a um "determinante" quando tanto ela [a propriedade] quanto o conjunto de propriedades contrárias a ela caem sob o determinante. Assim, o vermelho (junto com seus contrários verde, amarelo, etc.) é uma propriedade determinada em relação às cores; o escarlate é uma propriedade determinada em relação ao vermelho ou às cores. (MILLIKAN, 1984, pp. 20-1)

Considerando (c), portanto, podemos dizer que p_1 é a cor vermelha, p_2 é a cor verde e p_3 é a cor azul. Nesse contexto, (c) também é satisfeito, uma vez que a *lei in situ* descrita acima correlaciona três determinados (vermelho, verde e azul) a um determinante (cor) de modo que se alterássemos a parte vermelha do texto original para amarelo, uma nova cópia do documento também seria alterada.

No caso das dores, a sugestão é que analisemos dois estados mentais, um realizado em um sistema orgânico (um ser humano) e o outro em sistema composto por silício (um robô funcionalmente idêntico a nós). Dado que os robôs respondem do mesmo modo que nós quando estamos com dor, podemos identificar a existência de dois estados transitórios entre $inputs$ e $outputs$ em cada sistema. Dado ainda que a teoria epistêmica é verdadeira para casos biológicos, sabemos que no caso dos humanos esse estado transitório é uma dor. Também sabemos que toda realização atual desse estado é um reprodução (no sentido de Millikan) de um tipo M que foi selecionado no passado.

Com isso, podemos finalmente identificar a FOS a que todas realizações de dores nos seres humanos pertencem. De modo mais específico, todas essas instanciações da dor ($p_1 \ldots p_n$) pertencem a uma FOS P porque (i) $p_1 \ldots p_n$ são similares em certos aspectos (todos conectam $inputs$ e $outputs$ similares); (ii) essa similaridade é explicada por uma *lei in situ* (a universalidade das leis físicas que subjazem às leis biológicas); e (iii) as propriedades de reprodução estabelecidas $q_1 \ldots q_n$ (o comportamento associado a cada realização de dor) são explicadas por uma *lei in situ* (leis da física que subjazem aos processos biológicos).

Com essa caracterização, a questão sobre a múltipla realização pode ser posta do seguinte modo: pode um estado transitório m_h entre $inputs$ e $outputs$ em humanos ser da mesma natureza que um estado transitório m_r entre $inputs$ e $outputs$ de um robô? Ou, de modo mais

intuitivo, podemos dizer que robôs tem dores como seres humanos? Para que a teoria epistêmica se aplique a sistemas não-orgânicos, é preciso que a resposta seja sim.

É importante lembrar que a teoria epistêmica só pode afirmar isso quando os robôs em consideração tiverem uma história de seleção idêntica (ou muito similar) à história seletiva dos seres humanos. Isso quer dizer que o que nos permite dizer se eles sentem ou não dor é o conhecimento de sua história seletiva. Em outras palavras, uma mesma história seletiva implica similaridade funcional, o que, por sua vez, implica identidade de $qualia$ na perspectiva do teleofuncionalismo. Isso se torna claro na medida em que observamos que o estado transitório m_r no robô pertence a mesma família de reprodução estabelecida que o estado transitório m_h em seres humanos.

Uma dificuldade que surge ao sustentarmos essa identidade é a de especificar como sabemos que m_h e m_r pertencem à mesma FOS. A resposta, novamente, se encontra em um importante conceito desenvolvido por Millikan (1984): o conceito de *condições Normais* (com "N" maiúsculo). Para Millikan,

> Se algo x (a) foi produzido por algo que tem como função própria produzir um membro ou membros de uma família de reprodução estabelecida R, e (b) [x] é em alguns aspectos como membros Normais de R porque (c) foram produzidos de acordo com uma explicação que se aproxima em algum grau (indefinido) a uma explicação Normal para a produção de membros de R, então x é um membro de R. (MILLIKAN, 1984, p. 25)

De modo resumido, condições Normais são as condições nas quais ocorrências $o_1 \ldots o_n$ de um tipo T realizaram sua função historicamente. Assim, o que nos permite dizer que identidade de história seletiva implica identidade da realização de dores em sistemas de natureza distinta é o fato de que esses estados foram selecionados para realizar suas funções nas mesmas condições Normais. Isso torna claro, por outro lado, como a condição (b) estabelecida por Millikan é satisfeita no caso dos robôs, uma vez que m_r é igual a membros Normais ($p_1 \ldots p_n$) de membros da FOS P ao qual m_h pertence. Mais ainda, (c) também é satisfeito, já que a produção de m_r é explicada por explicações Normais que são similares àquelas associadas à produção de $p_1 \ldots p_n$.

Resta-nos compreender como a condição (a) é satisfeita. É nela que se situa o ponto crucial em favor da teoria epistêmica. Note que para satisfazer (a) é preciso que m_r tenha

sido produzido por um dispositivo que tem a função própria — ou etiológica — de produzir membros de P. Assim, é preciso que o robô tenha sido designado a produzir m_r em condições apropriadas. Isso quer dizer que para m_r ser membro de P, é necessário que o proceso seletivo que designa o robô a fazer m_r seja similar em seu aspecto lógico ao processo seletivo que permitiu seres humanos realizarem m_h.

Finalmente, podemos entender como um robô pode ter a mesma dor que temos. Vimos na seção 4.1 desse capítulo que, para Dretske (1995), para um estado A representar algo é preciso que A tenha sido *designado* a representar algo. No caso do robô, ele foi designado a representar as mesmas propriedades que nós representamos, já que m_h e m_r foram produzidos sob as mesmas condições Normais. Isso nos deixa com a conclusão de que o robô representa a mesma coisa que nós, e que se pretendemos falar dos $qualia$ de um ponto de vista evolutivo, então não há motivos para negar que robôs têm os mesmos $qualia$ que nós. Desse modo, formulamos a segunda versão da teoria epistêmica:

> (TEQ_2) Podemos justificar nossas crenças que outros seres sistemas possuem os mesmos $qualia$ que nós porque esses $qualia$ são membros de uma FOS, e como tal, são reproduções de um mesmo $quale$ ancestral que foram selecionados sob as mesmas condições Normais. Se houver uma diferença nos $qualia$, então deve haver também uma diferença na funcionalidade do material substrato, uma vez que o que tem uma função também deve ser capaz de mal-funcionar.

Essas considerações nos permitem concluir que no que diz respeito aos aspectos funcionais, não há um problema epistêmico dos $qualia$. Assim, a resposta que o teleofuncionalismo dá ao caso dos $qualia$ ausentes é que uma vez que certa organização funcional é realizada em conjunto com um processo de seleção adequado, a nossa crença sobre os $qualia$ de outras pessoas, animais, e possíveis sistemas com consciência pode ser justificada. Essa resposta, no entanto, não pretende sustentar que o sistema funcional composto por chineses tenha $qualia$, já que identidade funcional não é *suficiente* para identidade no nível mental. Dennett (1987), por exemplo, questiona se estaríamos dispostos a atribuir mente a sistemas funcionais idênticos a nós, mas que processassem informação de modo substancialmente mais lento. Embora não possamos tratar desse problema aqui, o que a teoria epistêmica afirma se alinha às preocupações de Dennett. Em outras palavras, embora identidade de organização funcional seja um requisito

necessário, ele não é suficiente para a presença de $qualia$. Para a teoria epistêmica, é preciso que haja identidade nos aspectos lógicos das histórias de seleção.

Nesse sentido, a teoria epistêmica sugere uma atitude cautelosa em relação a considerações que recorrem a dimensão espacial de um sistema funcional. Como mostra Lycan (1995), o cenário dos $qualia$ ausentes parece se sustentar em alguma tese chauvinista no que diz respeito ao tamanho daquilo que consideramos como consciente. Imagine que, após você dormir, alguns cientistas malignos o transforme num indivíduo tão pequeno, tal que você possa ver neurônios dentro do cérebro. Se, em seguida, os cientistas lhe disserem que você está dentro de um cérebro humano e que aquela pessoa está tendo uma experiência visual do vermelho, você provavelmente diria que essa é uma afirmação pouco intuitiva, uma vez que não há nada ali que seja vermelho. A teoria epistêmica é sensível a essas peculiaridades na medida em que ela não assume que o sistema funcional composto por chineses tenha mente, mas permite que, caso esse sistema tenha uma história seletiva similar a do ser humano, então ele poderia ter mente.

4.4 Mary e o morcego: problemas epistêmicos

A discussão até o momento esteve focada nos $qualia$ invertidos e nos $qualia$ ausentes. A questão que surge, no entanto, é como o teleofuncionalismo pretende lidar com as objeções levantadas por Nagel (1974) e Jackson (1982, 1986). Uma breve retomada desses pontos será útil, principalmente em relação à noção de $qualia$ que eles adotam.

Lembremos do caso de Mary. Quando dizemos que Mary descobre algo novo quando deixa seu quarto e olha para uma rosa, essa afirmação é sustentado pelas três características dos $qualia$ que discutimos no capítulo 3: isto é, que os $qualia$ (i) são propriedades intrínsecas das experiências; (ii) que são inefáveis ou subjetivos; e (iii) que são propriedades brutas ou monádicas. Assim, podemos dizer que Mary descobre algo novo quando ela vê um objeto vermelho porque (i) a "vermelhidão" é intrínseca à experiência do vermelho e, desse modo, só pode ser conhecida quando se tem essa experiência; (ii) a "vermelhidão" é uma propriedade inefável e subjetiva que não pode ser conhecida de modo objetivo; e (iii) porque a "vermelhidão" é uma propriedade bruta ou monádica que não pode ser conhecida em termos mais elementares.

Frente a esse cenário, o teleofuncionalismo afirma que os problemas decorrentes dessa

caracterização se restringem a um âmbito bastante específico. Lembre-se que a concepção de $qualia$ que sustenta o cenário hipotético de Mary é uma concepção dada no nível pessoal. Argumentei, no entanto, que essa concepção é muito restritiva, não devendo ser a única a informar o debate. Como alternativa, propus a adoção de teorias *interfield* que integram contribuições dadas por disciplinas que abordam a mente tanto de um nível pessoal (psicologia) quando de um nível sub-pessoal (neurociência). Um aspecto importante da proposta *interfield* de Darden e Maull (1977) é que problemas difíceis em um domínio podem ser resolvidos por experimentos e técnicas de outros domínios. Nesse sentido, se adotarmos a estratégia de decomposição em instituições, tal como sugere Lycan (1995), então uma análise a partir do nível sub-pessoal pode ajudar a resolver dificuldades no nível pessoal.

O que quero sugerir é que as dificuldades levantadas no caso de Mary podem ser dissolvidas quando adotamos uma análise de nível sub-pessoal. Em outras palavras, quando estudamos a mente a partir dos pressupostos de teorias *interfield*, não precisamos nos restringir a um único domínio de análise. Mais ainda, se pudermos mostrar que os problemas colocados tanto por Jackson quanto por Nagel são problemas restritos a um único domínio, então o teleofuncionalismo pode resolvê-los apelando para outros domínios sem violar seus pressupostos básicos. Desse modo, como o teleofuncionalismo não adota a concepção tradicional descrita no Capítulo 3 como absoluta, e como os argumentos de Nagel e Jackson dependem dessa concepção, os problemas levantados por eles não se aplicam ao teleofuncionalismo.

Antes de prosseguirmos de modo mais detalhado nesse ponto, considere o fato de que conhecer tudo sobre a vida de um morcego ou sobre o funcionamento do cérebro humano sem que tenhamos experiências relacionadas a esse conhecimento não implica, de modo necessário, um problema ontológico. Em outros palavras, o fato de conhecermos um manual de neurociência humana ou de morcegos sem sabermos como é ter as experiências ali descrita permite somente levantar problemas epistemológicos. Para ver isso de modo mais claro, considere o caso de John, um supercientista, que elaborou um manual perfeito descrevendo os processos físicos que ocorrem no mundo quando alguém anda de bicicleta. Sempre muito ocupado com seus estudos, John nunca aprendeu a andar de bicicleta, mas agora que terminou seu trabalho, resolve colocar seu conhecimento em prática. Não seria surpreendente se John, um especialista

na ciência de andar de bicicleta, sofresse alguns tombos nas primeiras vezes que tentasse andar de bicicleta. Isso indica que ainda que John saiba tudo sobre andar de bicicleta, ele ainda não sabe *como é* andar de bicicleta, o que parece bastante razoável.

O que esse caso pretende mostrar é que nos cenários colocados por Jackson e Nagel, tomamos como certo que as dificuldades levantadas por eles colocam um problema ontológico em cena, uma vez que se não podemos conhecer algo pelas suas propriedades físicas, então esse algo não é físico. O problema com esse movimento argumentativo é que teríamos que dizer há um domínio ontológico igualmente distinto no caso de John, o que parece absurdo. Isso se dá porque a experiência de John consiste apenas em um novo modo de conhecer um mesmo fato do mundo. No caso do manual, John representa seu conhecimento de modo proposicional, enquanto no segundo caso o representa de modo motor.

Essa distinção de modos de conhecimento é usualmente descrita em termos de *saber que* e *saber como*. Quando aplicamos essa terminologia ao caso de John, podemos dizer que antes de subir em sua bicicleta, John sabia *que* determinadas proposições eram o caso, ao passo que para andar em sua bicicleta, ele precisa saber *como* fazer isso. Do mesmo modo, podemos fazer a questão inversa: muitos de nós, que não somos cientistas, ao sermos questionados como andamos de bicicleta, simplesmente dizemos: "Não sei, eu apenas o faço". A questão que se coloca, portanto, é a seguinte: se não pretendemos dizer que há um domínio ontológico distinto no caso de John, por que deveríamos fazê-lo nos casos de Nagel e Jackson? Mais precisamente, não seriam esses problemas apenas de natureza epistêmica? Se esse for o caso, podemos concluir que os argumentos de Nagel e Jackson não apresentam bases sólidas para uma conclusão de natureza ontológica, explicitando apenas um problema epistemológico com o qual o teleofuncionalismo é capaz de lidar.

Para concluir essa seção, uma última consideração. David Lewis (1988) apresenta argumentos similares contra essas objeções, isto é, Lewis também acredita que o problema se situa no âmbito epistêmico. A minha tese, no entanto, se distingue da de Lewis em alguns pontos. Lewis sustenta que o fato de Mary não saber como é ver um objeto vermelho se deve a Mary não ter certa habilidade, isto é, a habilidade de conceber o vermelho, de imaginar o vermelho, de reconhecer o vermelho, etc.

A posição que defendo não se compromete com esse tipo de conclusão. Pretendo somente enfatizar que a distinção entre "saber que" e "saber como" explicita dois modos distintos de se representar um mesmo fenômeno. No caso de Mary, ela representa um mesmo fato do mundo de modos diferentes: no caso dos livros, de modo proposicional, e no caso da experiência visual, uma representação visual que não precisa ser necessariamente proposicional. Desse modo, não havendo distinções ontológicas, mas apenas no modo de representação, os argumentos de Jackson e Nagel caem no domínio da epistemologia. Isso quer dizer que se a teoria epistêmica junto ao teleofuncionalismo oferece boas razões para descartar esses problemas, então temos bons motivos para pelo menos adotar como hipótese de trabalho a ideia de que os $qualia$ podem ser explicados cientificamente, ainda que ao final descubramos que essa hipótese esteja equivocada.

4.5 Conclusão

Consideramos, ao longo desse trabalho, o que chamei *problema epistêmico dos qualia*. Esse problema coloca em foco o seguinte questionamento: como podemos saber se outros sistemas biológicos ou não-biológicos têm experiências conscientes com os mesmos aspectos qualitativos ($qualia$) que as nossas próprias experiências conscientes? Como ponto de partida, apresentei uma breve discussão sobre as principais teorias desenvolvidas ao longo do século XX para explicar o que é a mente. No contexto dessa discussão, argumentei que uma dessas teorias — o funcionalismo — se destaca frente às outras. Desse modo, sugeri que baseássemos nossa discussão a partir dessa teoria.

Embora se destaque frente a outras teorias, o funcionalismo também enfrenta sérios problemas. Discuti brevemente alguns desses problemas, optando por focar no problema sobre a possibilidade de inversão dos *qualia*. De modo breve, esse cenário questiona a plausibilidade do funcionalismo na medida em que parece possível inverter-se os $qualia$ de um estado mental sem alterar sua definição funcional. A possibilidade da inversão de espectro, aliada a possibilidade da ausência completa de *qualia* em sistemas não-biológicos, motivou a formulação do problema epistêmico. A minha suggestão para salvar o funcionalismo dessas dificuldades foi a de reformular a noção de função usada para definir os estados mentais. Ao invés de adotarmos

uma noção matemática de função, argumentei que seria mais frutífero conceber o funcionalismo a partir de uma noção etiológica de função, que preza pela origem evolutiva da função atribuída ao objeto de análise.

A reformulação do funcionalismo a partir da noção etiológica deu origem ao que chamei de teleofuncionalismo. Essa versão alterada do funcionalismo reconhece a origem histórica e seletiva da função dos estados mentais, permitindo conceber esses estados mentais como tendo um aspecto normativo, isto é, um estado mental só é de determinada natureza se realizar a função que foi designado (ou selecionado) a realizar. Assim, se um estado mental não realizar essa função, deve haver alguma diferença funcional ou causal no sistema, visto que essa é uma condição necessária para podermos dizer que algo realiza sua função de modo correto ou incorreto. Desse modo, quando aplicado ao caso de estados mentais com $qualia$, podemos dizer que, de acordo com o teleofuncionalismo, se houver inversão de $qualia$, deve haver necessariamente diferenças funcionais no sistema.

O reconhecimento do aspecto normativo das definições teleofuncionais dos estados mentais nos permite, finalmente, formular a teoria epistêmica dos *qualia*. Essa teoria, em sua primeira formulação, diz que sabemos que outros seres humanos possuem estados mentais com *qualia* porque nossos estados mentais ou conscientes são reproduções de um estado mental ancestral com *qualia* que foi selecionado em certas condições Normais. Do mesmo modo, na segunda formulação da teoria, podemos dizer que sistemas não-biológicos possuem estados mentais com *qualia* como os nossos se seus estados mentais ou conscientes foram selecionados por (i) um processo de seleção da mesma natureza que a seleção natural e (ii) esses estados foram selecionados sob as mesmas condições Normais. Desse modo, dado que estados mentais são membros de Família de Ordem Superior (FOS), sabemos que os estados mentais com *qualia* que temos hoje são reproduções de estados mentais selecionados no passado, e que se esses estados mentais sofererem alguma alteração, como a inversão de espectro, deverá haver necessariamente uma diferença funcional no sistema. Sendo esse o caso, podemos então conceber uma versão do funcionalismo que oferece uma resposta consistene ao problema epistêmico dos $qualia$.

Referências

AKINS, K. What is it like to boring and myopic?. In: DAHLBOM, D. *Dennett and His Critics*. Cambridge: Blackwell, 1993.

BLOCK, N. Troubles with Functionalism. In: BLOCK, N. (Org.). *Readings in the Philosophy of Psychology, Volume 1*. Cambridge: Harvard University Press, 1980.

____. On a confusion about the function of consciousness. *Behavioral and Brain Sciences*, n. 18, p. 227-247, 1995.

BROGAARD, B. Introduction: Does Perception Have Content?. In: BROGAARD, B. (Org.). *Does Perception Have Content?*. Oxford: Oxford University Press, 2014.

BULLER, D. Etiological Theories of Function: A Geographical Survey. *Biology and Philosophy*, v. 13, p. 505-527, 1998.

CHALMERS, D. *The Conscious Mind*. New York: Oxford University Press, 1996.

CHURCHLAND, P. M. Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, v. 78, n. 2, p. 67-90, 1981.

____. *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind*. Cambridge: MIT Press, 1984.

____. (1989). Knowing Qualia: A Reply to Jackson. In: CHURCHLAND, P. M; CHURCHLAND, P. S. *On the Contrary*. Cambridge: MIT Press, 1996.

CHURCHLAND, P. M; CHURCHLAND, P. S. *On the Contrary*. Cambridge: MIT Press, 1996.

COHEN, M.; DENNETT, D. Conscioussness cannot be separated from function. *Trends in Cognitive Sciences*, v. 15, n. 8, p. 358-364, 2011.

CRANE, T. Is there a perceptual relation?. In: GENDLER, T.; HAWTHORNE, J. (Orgs.). *Perceptual Experience*. Oxford: Oxford University Press, 2006.

CUMMINS, R. Functional Analysis. *Journal of Philosophy*, v. 72, n. 20, p. 741-765, 1975.

DARDEN, L.; MAULL, N. (1977). Interfield theories. In: DARDEN, L. (Org.). *Reasoning in*

Biological Discoveries. Cambridge: Cambridge University Press, 2006.

DARDEN, L.; CAIN, J. (1989). Selection Type Theories. In: DARDEN, L. (Org.). *Reasoning in Biological Discoveries*. Cambridge: Cambridge University Press, 2006.

DENNETT, D. *Content and Consciousness*. New York: Routledge, 1969.

_____. *Brainstorms*. Cambridge: MIT Press, 1981.

_____. Three kinds of intentional psychology. In: DENNETT, D. *The Intentional Stance*. Cambridge: MIT Press, 1987.

_____. (1988). Quining Qualia. In: CHALMERS, D. (Org.). *Philosophy of Mind Classical and Contemporary Readings*. New York: Oxford University Press, 2002.

_____. *Consciousness Explained*. London: Penguin Books, 1991.

_____. *Darwin's Dangerous Idea*. New York: Simon and Schuster, 1995.

_____. Facing backwards the problem of consciousness. *Journal of Consciousness Studies*, v. 3, n. 1 , p. 4-6, 1996.

DESCARTES, R. (1641). *Meditações*. Trad. J. Guinsburg e Bento Prado Júnior. São Paulo: Nova Cultural, 1996. (Coleção Os Pensadores).

DRETSKE, F. *Naturalizing the mind*. Cambridge: MIT Press, 1995.

GODFREY-SMITH, P. Functions: Consensus Without Unity. *Pacific Philosophical Quarterly*, v. 74, n. 3, p. 196-208, 1993.

GRIFFITHS, P. Functional analysis and proper functions. *British Journal of the Philosophy of Science*, v. 44, n. 3, p. 409-422, 1993.

HARMAN, G. The intrinsic quality of experience. *Philosophical Perspectives*, v. 4, p. 31-52, 1990.

JACKSON, F. Epiphenomenal qualia. *Philosophical Quarterly*, v. 32, p. 127-136, 1982.

JAMES, W. (1890). *The Principles of Psychology*. Cambridge: Harvard University Press, 1981.

KANT, I. *Critique of Pure Reason*. Cambridge: Cambridge University Press, 1998. (The Cambridge Edition of the Works of Immanuel Kant).

KIM, J. *Philosophy of Mind*. Boulder: Westview Press, 1998.

KITCHER, P. (1993). Function and Design. In: RUSE, M. and HULL, D. (Org.). *The Philosophy of Biology*. Oxford: Oxford University Press, 1998.

KRIPKE, S. *Naming and Necessity*. Oxford: Basil Blackwell, 1980.

LEWIS, D. (1988). What experience teaches. In: CHALMERS, D. (Org.). *Philosophy of Mind Classical and Contemporary Readings*. New York: Oxford University Press, 2002.

LUCRÉCIO. *Da Natureza*. In: Coleção Os Pensadores. São Paulo: Abril Cultural, 1973.

LYCAN, W. Form, function, and feel. *The Journal of Philosophy*, v. 78, p. 24-50, 1981.

____. *Consciousness*. Cambridge: MIT Press, 1995.

MATTHEN, M. Our knowledge of colour. *Canadian Journal of Philosophy*, v. 27, n. 1, p. 215-46, 2001.

MCCAULEY, R.; BECHTEL, W. Explanatory Pluralism and Heuristic Identity Theory. *Theory and Psychology*, v. 11, n. 6, p. 736-760, 2001.

MILLIKAN, R. *Language, Thought and Other Biological Categories*. Cambridge: MIT Press, 1984.

____. An Ambiguity in the Notion "Function". *Biology and Philosophy*, v. 4, p. 172-176, 1989a.

____. In Defense of Proper Functions. *Philosophy of Science*, v. 56, n. 2, p. 288-303, 1989b.

____. Wings, Spoons, Pills and Quills: a Pluralist Theory of Functions. *Journal of Philosophy*, v. 96, n. 4, p. 191-206, 1999.

____. Biofunctions. In: ARIEW, A (Org.). *Functions*. Oxford: Oxford University Press, 2002.

MUNDALE, J.; BECHTEL, W. Integrating Neuroscience, Psychology, and Evolutionary Biology Through a Teleological Conception of Function. *Minds and Machines*, v. 6, p. 481-505, 1996.

NAGEL, T. What is it like to be a bat?. *Philosophical Review*, v. 83, n. 4, p. 435-456, 1974.

NEANDER, K. The Teleological Notion of Function. *Australasian Journal of Philosophy*, v. 69, n. 4, p. 454-68, 1991a.

____. Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of Science*, v. 58, n. 2, p. 168-184, 1991b.

NUDDS, M. Recent Work in Perception: Naive Realism and its Opponents. *Analysis Reviews*, v. 69, n. 2, p. 334-346, 2009.

O'CALLAGHAN, C. Lessons from Beyond Vision (Sounds and Audition). *Philosophical Studies*, v. 153, n. 1, p. 143-160, 2011.

_____. The Multisensory Character of Perception. *The Journal of Philosophy*, v. 112, n. 10, p. 551-569, 2015.

PAPINEAU, D. *Philosophical Naturalism*. Cambridge: Blackwell Publishers, 1993.

PLACE, U. T. Is Consciousness a Brain Process?. In: CHALMERS, D. (Org.). *Philosophy of Mind Classical and Contemporary Readings*. New York: Oxford University Press, 2002.

PUTNAM, H. *Brains and Behavior*. In: CAPITAN, W. H.; MERRILL, D.D. (Orgs.). *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, 1967.

PUTNAM, H. (1973). The Nature of Mental States. CHALMERS, D. (Org.). *Philosophy of Mind Classical and Contemporary Readings*. New York: Oxford University Press, 2002.

RORTY, R. *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press, 1979.

SANT'ANNA, A. The role of selection in functional explanations. *Manuscrito*, v. 37, n. 2, p. 227-267, 2014.

SEARLE, J. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, v. 3, 1980.

_____. *Intentionality*. Cambridge: Cambridge University Press, 1983.

_____. *The Rediscovery of the Mind*. Cambridge: MIT Press, 1992.

_____. *Mind: A Brief Introduction*. New York: Oxford University Press, 2004.

SMART, J. Sensations and Brain Processes. In: CHALMERS, D. (Org.). *Philosophy of Mind Classical and Contemporary Readings*. New York: Oxford University Press, 2002.

TYE, M. *Ten Problems of Consciousness*. Cambridge: MIT Press, 1995.

WIMSATT, W. Teleology and the logical structure of function statements. In: *Studies in History and Philosophy of Science*, v.3, n. 1, p. 1-80, 1972.

WRIGHT, L. Functions. *Philosophical Review*, v. 82, n. 2, p. 139-169, 1973.